DOCUMENT RESUME

ED 162 528 FL 009 894

TITLE Teaching English as a Second Language: Perspectives and Practices. A Series of Six Texts. Testing: Fourth of a Series.
INSTITUTION New York State Education Dept., Albany.
PUB DATE 78
GRANT G0077C0041
NOTE 89p.
AVAILABLE FROM The University of the State of New York, The State Education Department, Bureau of Bilingual Education, Albany, New York 12234

EDRS PRICE MF-$0.83 HC-$4.67 Plus Postage.
DESCRIPTORS Achievement Tests; *Cloze Procedure; *Communicative Competence (Languages); Culture Free Tests; *English (Second Language); Intelligence Tests; Language Dominance; Language Instruction; *Language Proficiency; *Language Tests; Learning Processes; Minority Group Children; *Second Language Learning; Sociolinguistics; Teaching Methods; Test Construction; Testing; Test Interpretation; Test Reliability

ABSTRACT

This book is designed to assist those who work with non-English dominant students by providing resource information relevant to second language teaching and learning. The articles in the series encompass both theory and practical learning techniques in six general topics. Five articles concerning testing are presented in the fourth volume of this series. "Testing," by Robert Cooper, is a brief introductory guide to the construction, selection, and use of second language tests, specifically tests of English as a second language. The second article, "Writing Proficiency and Achievement Tests," by Sydney Sako, discusses test items designed to measure the four types of language skills: listening comprehension, reading comprehension, speaking production, writing production, and combinations of these skills. The third article, "The Assessment of Language Dominance and Proficiency," by Marina Burt and Heidi Dulay, defines the dimension of bilingual measurement in order to assess oral language dominance and proficiency. The fourth article, "Using Cloze Procedure as an Overall Language Proficiency Test," by Kenneth Aitken, discusses the construction, administration, scoring and interpretation of cloze tests of overall language proficiency. The last article, "I.Q. Tests for Minority Children and a Piagetian-Based Computerized Information System as an Alternative," by Edward A. DeAvila, discusses the inadequacy of the response of test publishers to public criticism of I.Q. tests. Appended are a list of abbreviations and definitions used in the book and a list of materials and services available from the Bureau of Bilingual Education, New York State Education Department. (NCB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED162528

TEACHING ENGLISH AS A SECOND LANGUAGE: PERSPECTIVES AND PRACTICES

A Series of Six Texts

Testing

FOURTH OF A SERIES

The work upon which this publication is based was performed pursuant to ESEA Title VII Contract No. G0077C0041 of the United States Office of Education to the New York State Education Department. It does not, however, necessarily reflect the views of that agency or the New York State Education Department.

A limited number of copies are available upon request from:

The University of the State of New York
THE STATE EDUCATION DEPARTMENT
Bureau of Bilingual Education
Albany, New York 12234

1978

FL009894

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
New York State Education Dept
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

THE UNIVERSITY OF THE STATE OF NEW YORK
Regents of The University (with years when terms expire)

1981	Theodore M. Black, A.B., Litt.D., LL.D. Pd.D., D.C.L., L.H.D. Chancellor	Sands Point
1988	Willard A. Genrich, LL.B., L.H.D., LL.D. Vice Chancellor	Buffalo
1986	Kenneth B. Clark, A.B., M.S., Ph.D., LL.D., L.H.D. D.Sc.	Hastings on Hudson
1983	Harold E. Newcomb, B.A.	Owego
1982	Emlyn I. Griffith, A.B., J.D.	Rome
1983	Mary Alice Kendall, B.S.	Irondequoit
1984	Jorge L. Batista, B.A., J.D., LL.D.	Bronx
1982	Louis E. Yavner, LL.B.	New York
1979	Laura B. Chodos, B.A., M.A.	Clifton Park
1980	Martin C. Barell, B.A., I.A., LL.B.	Kings Point
1981	Joseph R. Bongiorno, B.S., M.D.	Brooklyn
1984	Louise P. Matteoni, B.A., M.A., Ph.D.	Bayside
1981	J. Edward Meyer, B.A., LL.B.	Armonk
1985	Arlene B. Reed-Delaney, B.A., M.D.	Albany
1987	R. Carlos Carballada, B.S.	Arcade

President of The University and Commissioner of Education
GORDON M. AMBACH

Executive Deputy Commissioner of Education
JOSEPH J. BLANEY

Deputy Commissioner for Elementary, Secondary and Continuing Education
ROBERT R. SPILLANE

Associate Commissioner for Instructional Services
VIVIENNE ANDERSON

Assistant Commissioner for General Education and Curricular Services
MARIA RAMIREZ

Director, Division for Curriculum Services
GORDON E. VAN HOOFT

Chief, Bureau of Bilingual Education
______________ ____________

FOREWORD

Teaching English as a Second Language: Perspectives and Practices is designed to assist those who work with non-English dominant students by providing resource information relevant to the processes of second language teaching and learning. The articles encompass both theory and practical learning techniques in six general topics and have been compiled in response to increased interest in the education of learners whose native language is other than English. It is becoming more widely recognized that such students can benefit from formal instruction in English as a second language as well as efforts to maintain and utilize their native language and culture. This series replaces the previous *Teaching English as a Second Language Series*.

This is the fourth of a series of six volumes. Other volumes of the series are the following:

Background and Approaches
Speaking and Understanding
Reading and Writing
Current Issues in Second Language Teaching
English as a Second Language and Bilingual Education

Acknowledgments

The work of the consulting editor, Dr. Richard Light, Associate Professor in Teaching English as a Second Language, State University of New York at Albany, is appreciated. Laurie Wellman, Associate, Bureau of Bilingual Education, organized and prepared the manuscript for publication. The Department is grateful to all who granted permission to include their articles in this publication.

TEACHING ENGLISH AS A SECOND LANGUAGE
PERSPECTIVES AND PRACTICES

A Series of Six Texts

Contents

Testing:
Fourth of a Series

*1. *Testing*
Robert Cooper-- 1

*2. *Writing Proficiency and Achievement Tests*
Sydney Sako---16

3. *The Assessment of Language Dominance and Proficiency*
Marina Burt and Heidi Dulay---------------------------29

4. *Using Cloze Procedure as an Overall Language Proficiency Test*
Kenneth Aitken--52

*5. *I.Q. Tests for Minority Children and a Piagetian-Based Computerized Information System as an Alternative*
Edward A. DeAvila-------------------------------------67

Additional Information:

1. List of Abbreviations and Definitions Used in this Text----80

2. Materials and Services Available from the Bureau of Bilingual Education, New York State Education Department---82

---

* Asterisk indicates article is retained from first edition of this series. Other articles are appearing for the first time. Articles with asterisk are single-spaced; others double spaced.

TESTING

Robert L. Cooper

There are many users and uses of second-language tests. Admissions officers at American universities, for example, rely on scores obtained from tests of English as a second language to help evaluate the qualifications of foreign applicants. Directors of programs which teach English as a second language employ tests to help place students in classes at an appropriate level of instruction. Teachers of these classes use tests to help them to determine whether or not students have mastered a given unit of instruction or to identify students' strengths and weaknesses or to rank students for purposes of assigning grades. Teacher trainers use tests to help evaluate the effectiveness of alternative methods of instruction. Language tests, in short, are used at all stages of second-language instruction: for the admission and placement of students, for the evaluation of the success with which instructional objectives are met, and for the prediction of the success with which language learners can use the second language both inside and outside the classroom.

It is the purpose of this chapter to provide a brief introductory guide to the construction, selection, and use of second-language tests, with particular reference to tests of English as a second language. For more detailed treatments, the reader is referred to Lado (1961) and to Valette (1967) for second-language testing generally and to Harris (1969) for the testing of English as a second language specifically. Useful discussions of the problems involved in second language testing may also be found in Pimsleur (1966), Davies (1968), and Upshur and Fata (1968). For an exposition of testing issues more generally, the reader is referred to Thorndike and Hagen (1969). Reviews of published tests of English as a second language may be found in Buros (1965).

Philosophical Preliminaries

For our purposes, we may define a test as a task or set of tasks which can be assigned under uniform conditions for purposes of assessment or evaluation. When we ask a student to respond to a question, for example, we have assigned him a task. If we can ask the same question in the same way to other students (or to the same student at a later time), then we are able to assign the task under uniform conditions. And if the task is assigned primarily to collect information about the people who perform it, then the task may be considered a test. We are not concerned here with the informal evaluations which language teachers make in response to their students' performance on relatively uniform tasks such as pattern-practice exercises or written assignments. We are concerned instead with those tasks which are assigned primarily for information gathering purposes and which are relatively formal in nature, i.e., they are customarily viewed as tests both by those who administer them and by those who take them.

There is little point in asking people to take our tests if the resulting test scores are not put to use. Therefore, before we seek, construct, or administer a test, we must ask ourselves what use we expect to make of

---

Robert Cooper has taught at Yeshiva University, New York, New York 10027.

test scores. Before we use a test, in other words, we must ask ourselves if the information which it provides is relevant to any decision we must make about the individuals who take the test. All too often tests are administered routinely, the original purpose of testing forgotten, and the scores serving no purpose beyond that of employing testmakers, examiners, scores, and file clerks. Are we to give a test of English listening comprehension, for example? Then we must ask why we need to know about the examinees' English listening comprehension ability. We would be more justified in trying to obtain this information if we were to use the test scores to evaluate our success in teaching English listening comprehension, or to decide whether the examinees are likely to be able to understand lectures in English, than if we planned to use the scores to decide whether they could read technical material in English well enough to pursue graduate study. We would certainly not be justified in administering the test, however, if the scores were not consulted at all. Testing, in short, is not justified unless decisions are made on the basis of test scores and unless the information provided by the test scores is relevant to these decisions.

Not only must we ask if the test scores provide relevant information which will be used, we must also ask if the information we seek cannot be obtained in other ways. Testing is only one way to gather information about people. If we wanted to make judgements about a student's ability to understand spoken English, for example, we could play a taped conversation to him and then ask him to answer questions designed to assess his comprehension of the passage. But we could also gather information about his listening comprehension in English by asking him to tell us how well he can understand English. Or we could ask people who have had an opportunity to observe him, in situations where English is spoken, to rate his ability. Or we could rate him ourselves on the basis of our own past observations of him. Or we could, perhaps, make inferences about his ability from the grades he has received in courses in which English was the medium or the subject of instruction. Each of these means of gathering information about the student has certain advantages and disadvantages, and the test does not necessarily have more advantages on balance than the others. Tests do offer certain advantages: they facilitate ranking the performances of individuals, since the individuals are presented with identical tasks, and they usually offer greater precision of measurement. On the other hand, they are often expensive to develop, and they require time and money to administer and score, and they typically elicit only a tiny sample of the behavior that we want to evaluate. Thus, the decision to use a test ought to be made in terms of the usefulness and cost of the information which it yields in comparison to the usefulness and cost of information available from other sources.

The Content of Language Tests

Once we have decided to gather information by testing, we must specify the content of the test. That is to say, we must describe the kinds of behavior to be elicited by the test. What kinds of performances should a test elicit, for example, if we want to make valid inferences about a student's control of English grammar? What kinds of test behaviors will allow us to make valid inferences about a student's ability to read university-level textbooks written in English? What should a test contain if it is to yield scores which are maximally predictive of an elementary-school student's ability to speak English in a classroom setting? We are unlikely to obtain tests which will permit sound inferences unless we first specify the tests' content.

There are some skills for which it is relatively simple to specify test content. For example, if we want to know if a person can ride a bicycle, we can give him a bicycle and ask him to ride it. Similarly, if we want to know if a person can swim 20 yards, we can bring him to a pool 20 yards long and ask him to swim its length. In these examples, the skill in which we are interested and the test which allows us to measure it are identical. When we wish to test proficiency in English, however, our task is not so simple. Unlike bicycle riding or swimming, English proficiency is not identical to any task which is used to assess it. We can learn whether or not a man can ride a bicycle or swim 20 yards by giving him a test. We can, it is true, learn something about a man's English proficiency by means of test scores, but the conclusions we draw from them are far more tentative, since we can only observe a small fraction of all the behaviors which English proficiency makes possible.

There are really two problems involved in specifying the content of a language test. First, we must determine what it is we wish to measure. Second, we must specify the tasks which allow us to make valid inferences about what we want to measure. Do we want to measure English proficiency? Then first we must decide what we mean by English proficiency. Do we mean implicit knowledge of English syntax? Do we mean knowledge of a basic vocabulary? Do we mean the ability to read English at a given level of difficulty? Once we have decided *what* we want to measure, then we can decide *how* to go about measuring it.

We have identified two steps in the specification of test content: (1) determining *what* it is we want to test and (2) determining *how* to test it. Let us consider each of these steps in turn.

Determining What To Test

The determination of what it is we want to test involves making specifications along each of three dimensions: language variety, knowledge, and skill. "Language variety" refers to the dialect, register, style, or level of formality in which proficiency is to be tested. "Knowledge" refers to linguistic competence in the selected variety or varieties; and "skill" refers to the behaviors through which this knowledge or competence is realized, specifically listening, speaking, reading, or writing.

Variety

Recent work on sociolinguistics has emphasized the systematic variation which appears to exist in the use of different language varieties by members of a given speech community. Systematic alternation between varieties is more obvious where the varieties are viewed as separate languages and where the speech community is thought of as "bilingual" (e.g. ,the use of Spanish and English by the Puerto Rican speech community in New York City) than where the varieties are viewed as different styles or levels of formality of the same language and where the speech community is thought of as "monolingual." Even "monolingual" speakers, however, have more than one style of speech. Consider, for example, an American university professor's use of English. His style of speech will not be the same when he talks to his wife at home as when he talks to the president of his university when in the latter's office; similarly, the way he delivers a paper before a plenary session of a learned society will differ from the way he talks to a close friend when both are seated at a prizefight. The variations in his speech which are systematically associated with the different social contexts in which he enters can be described explicitly in terms of specific phonological, morphophonemic, syntactic, and lexical variants, and it is one of the tasks of sociolinguists to do this. Even if we were unable to describe explicitly the nature of the variation, however, we recognize implicitly, on the basis of our everyday experience, the existence of such variation.

Thus, when we define the content of an English proficiency test we must identify the variety or varieties of English that the test should represent. The content of a test designed to assess the ability of a student to understand the kind of English used in formal university lectures, for example, ought to differ from the content of a test designed to assess the ability of a student to understand the informal conversation of fellow students outside of class. The varieties to be tested could be identified in terms of relatively intuitive notions such as "formal standard," "casual standard," or "intimate standard," or in terms of the contexts in which English is to be used, e.g., university seminar, professional consultation, relaxed peer-group conversation, etc. In either case, recognition should be given to the code switching which typically occurs among so-called "monolingual" speakers of English. It would be appropriate to frame test content in terms of the speaker's knowledge of only a single code if, and only if, we are interested in measuring his knowledge of one code only.

Knowledge

Once we have identified the variety or varieties of English that we wish to cover, we must select the particular aspect or aspects of competence that we wish to test. If we accept the generative transformational grammarians' view of a language (or variety) as an infinite set of sentences, then linguistic competence is the ability to produce and understand the infinite set (Chomsky 1965). The speaker-hearer's competence can be viewed as that knowledge which enables him to produce any and all the sentences in this set. It is the task of grammarians to specify or make explicit this knowledge. According to the formulations of the generative transformational grammarians, there appear to be three components of this knowledge: (1) syntactic, (2) phonological, and (3) semantic.

Syntax

The syntactic component enables speakers to assign structural descriptions to sentences. That is, knowledge of the syntactic component enables us to distinguish grammatical English sentences, e.g., *the boy is* *hitting the ball,* from nongrammatical ones, e.g., *The is ball hitting by the.* Knowledge of the syntactic component also enables speakers to identify two sentences with identical structures, e.g., speakers would agree that the sentences *The boy is here* and *The girl is here* are structurally identical (although lexically different), whereas they would also agree that the sentence *The boy is here* and *Is the boy here* are not structurally identical. Knowledge of syntax also enables speakers to identify syntactically ambiguous sentences, so that speakers would agree, for example, that the sentence *He likes entertaining girls* has two syntactic interpretations.

Knowledge of the phonological component enables speakers to assign phonetic interpretations to sentences. That is, phonological competence allows speakers to produce and attend to those phonetic features of the stream of speech which are semantically empty but criterial. So for example in the sentence *The ship is here,* speakers of English will interpret the vowel in *ship* as /I/. Let us imagine, however, a speaker who has perfect command of English syntax but no knowledge of English phonology (e.g. a speaker of Spanish who has learned English from reading but has never heard it spoken). He would be able to assign a syntactic description to this sentence, e.g., he would know the sentences *The ship is here* and *The sheep is here* were structurally identical. However, he would not be able to distinguish the two sentences without contextual clues because he would not attend to the difference between /I/ and /i/ which is criterial for speakers of English but not for speakers of Spanish.

Semantics

The semantic system enables speakers to interpret sentences. If a speaker knew only the syntax of English he would know that the two sentences *The boy is here* and *The girl is here* were structurally identical. However, he would not know that the two sentences differed in meaning. Knowledge of the semantic component enables speakers to give different semantic interpretations to sentences which are structurally identical. Conversely, knowledge of semantics enables speakers to assign the same interpretation to sentences which are structurally different, e.g., *The boy hit the ball* and *The ball was hit by the boy.* Also, knowledge of the semantic component enables speakers to identify anomalies. For example, speakers of English would identify as anomalous Chomsky's (1957) famous sentence *Colorless green ideas sleep furiously,* or such sequences as *The table laughed* or *The dainty elephant.* In addition, knowledge of semantics enables the speaker to identify semantically ambiguous sentences. Thus, for example, he could assign two meanings to the sentence *We saw the late Mr. Smith.*

We may be interested in assessing the examinee's knowledge of one or more of the three components individually. For example, we may want to assess the degree to which he understands a given technical vocabulary, or we may want to measure the degree to which he is able to interpret syntactically complex spoken sentences. We may wish to make assessments in terms of individual components of competence because they represent specific goals of instruction or for other reasons. Thus, test content can be specified in terms of the degree to which each of the components of linguistic competence-phonological, syntactic, or semantic-should be covered specifically.

We may not, however, be particularly interested in the degree to which the examinee has mastered a given component of linguistic competence. We may, instead, be interested in the degree to which the examinee can combine all of these components into an integrated performance (Carroll 1961). If, for example, our purpose is to rank the examinees with respect to their ability to understand spoken English, we need not concern ourselves with their knowledge of any specific component comprising that skill. Instead, we can ask them to interpret recorded conversations, for example, a task which requires the simultaneous operation of phonological syntactic, and semantic knowledge. Thus, in addition to specifying the various components of competence which we may want to test, we must indicate whether or not we wish to assess the ability to combine all of these components simultaneously.

Skill

In principle, a speaker-hearer's knowledge of his language is the same for encoding as for decoding sentences. That is, the grammars which are written to account for his ability to produce the sentences of his language are also considered to account for his ability to understand the sentences of his language. However, his knowledge is realized through the application of different skills: speaking, listening, reading, and writing. People who speak the same first language differ with respect to their abilities as speakers, or as listeners, or as readers, or as writers of that language, even if they can all be presumed to have the same underlying linguistic competence. Just as there are interpersonal differences, there are intrapersonal differences. One person may be better at reading, for example, than at writing. Similarly, second-language learners may exhibit interpersonal as well as intrapersonal differences in the skills through which their knowledge of the second language is manifested. Thus, in addition to specifying the variety of English and the aspects of knowledge which are to be tested, we must specify which skills we want to assess-listening, speaking, reading, or writing.

Language Testing Framework

The content of a language test, therefore, can be specified in terms of a three-dimensional framework. For illustrative purposes, let us assume that we have identified two varieties of English, knowledge of which we want to assess, and that we have labelled these varieties as A and B which may, for example, correspond to "formal" and "informal" standard English respectively. Our framework will then appear as in figure 1. In this figure we have specified four skills (listening, speaking, reading, and writing) on the vertical axis, four aspects of knowledge (phonology, syntax, and semantics plus a combination of all three aspects or "total") on the horizontal axis, and two varieties (A and B) on the diagonal axis. The framework yields 4 X 4 X 2 = 32 theoretically distinct "cubes," each formed by the intersection of a particular skill, aspect of knowledge, and variety of English. For example, the cube formed by the intersection of listening, semantics, and variety A would represent the ability of the examinee to assign semantic interpretation to spoken sentences in variety A; the cube formed by the intersection of writing, phonology, and variety A would represent his ability to spell in variety A; and the cube formed by the intersection of speaking, total and variety B would represent his ability to speak in variety B, employing simultaneously his knowledge of the phonological, syntactic, and semantic features of that variety. Note that in this framework the phonological aspect represents orthographic skills when viewed in terms of reading and writing.

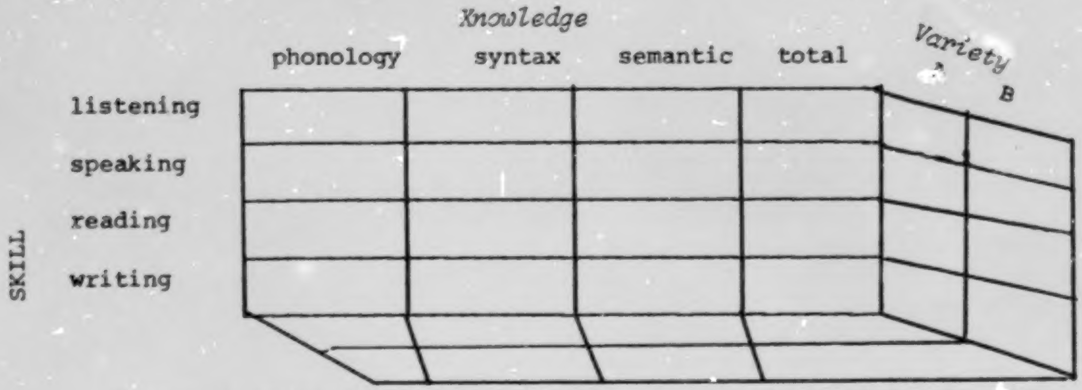


Figure 1: Framework for Specifying Language Test Content

The 32 cubes formed by our three-dimensional framework represent logical possibilities only. It is unlikely that we would want to find subtests or items to test each of these individually. We may, for example, be interested in only one skill. Or we may be interested only in knowledge when all aspects are manifested simultaneously in an integrated (total) performance. Furthermore, the ability represented by any given cube is likely to covary with the ability represented by any other. Thus, for example, although interpersonal differences exist in the ability to speak or to listen or to read or to write a given variety, in general those who attain high scores on a test of speaking that variety are likely to have relatively high scores on a listening comprehension test of that variety. Similarly, for a given variety, knowledge of phonology is likely to vary with knowledge of syntax and semantics since each is an interrelated component of the same system. Similarly, knowledge of one variety is likely to vary with knowledge of the other, since their systems or grammars are related.

The decision as to which of the 32 cubes should be assessed depends upon the use which we plan to make of the test scores. For example, if our instructional objectives have been stated in terms of auditory and spoken control of variety B's syntax, and if we want to assess the degree to which these objectives have been met, we will assess the abilities represented by the two cubes (listening-syntax-variety B and speaking-syntax-variety B). If, on the other hand, we want to predict the degree to which students will be able to understand spoken variety A, we will test the abilities represented by only one cube (listening-total-variety A). In general, when we want to measure an integrated skill (the cubes involving "total" knowledge), we get at least as good results when we assess complex performances directly, i.e., performances involving the simultaneous operation of phonological, syntactic, and semantic knowledge, as we do when we combine the results obtained from subtests or items testing each of these components separately (Carroll 1968).

Determining the Tasks

Once we have decided what we want to assess in terms of particular language varieties, skills, and aspects of competence, we are faced with the problem of specifying the tasks to be included in the test. The problem of specifying the tasks is essentially a problem of sampling, since there are many more items through which it is possible to assess the abilities represented by a given cube than can be used in a single test. For example, let us suppose we want to assess the ability to make syntactic interpretations of sentences spoken in variety A (the cube represented by the intersection of listening comprehension, syntax, and variety A). There are several types of operation which could be employed to test this. For example, the examinee could be asked to listen to a series of utterances and to (1) identify those utterances which were well formed or, conversely, those which were structurally deviant, (2) correct those utterances which were structurally deviant, (3) identify pairs of utterances with identical structures, (4) identify a unit in one sentence which had a given analogous function in another, (5) identify utterances which were syntactically ambiguous, (6) supply alternative interpretations to syntactically ambiguous utterances, or (7) supply items which had been randomly deleted from utterances. No doubt other operations could also be specified to test the ability to assign syntactic interpretations to spoken sentences. The particular operations which are chosen for the test will depend in part on their suitability for examinees at a given age and at a given level of English proficiency, in part on the difficulty of writing items to embody such operations, and in part on the limitations imposed by time and money.

Not only do we have to determine the operations to be embodied in the tasks to be included, we also have to determine the linguistic or contextual content of the tasks. In the example of testing the ability to assign syntactic interpretations to spoken sentences, what sentences are we to choose? Since the number of possible sentences is in principle infinite, we can present the examinee with only an infinitesimally small sample of these sentences. Clearly, we must choose sentences which exemplify certain structural relationships. But here again the number of possible types is very large. We are faced with a similar problem when we want to test the integrated or "total" skills. Here we must select not individual sentences or sentence types exemplifying specified syntactic relationships, or individual lexical items, or particular phonological contrasts. Instead, we must choose passages of connected discourse. And the passages must be representative of specified communicative contexts. Thus, for example, if we wanted to assess a person's ability to understand variety B when spoken, we must choose passages representative of the kinds of situations in which he is likely to hear variety B. If, for example, we wanted to assess the ability of foreign graduate students to read English-medium texts in their field of study, we would choose one type of passage. On the other hand, if we wanted to assess the ability of Mexican-American primary-school children to read third-grade English-medium primers, we would choose another. The particular structural or contextual content chosen for the test will be selected in terms of the same considerations involved in choosing the operations to be included: suitability for the examinees, ease of writing, and cost of administration.

Once we have specified the content of the test in terms of particular cubes of the test specification framework and once we have decided the type of operations to be elicited and the kinds of linguistic and contextual content to be covered, we are ready to write the items for the test, or, if we are to select tests from published ones, to determine the extent to which the content of alternative tests is similar to the test content we have specified.

Steps in Test Construction

The degree of care employed and the amount of time and money spent to construct a test can vary greatly, depending upon the purpose of the test. A classroom teacher, for example, who wishes to assess the degree to which the objectives of a one-week unit of instruction have been met with ten students will clearly need to expend less time and effort in constructing a test than the publishers of a test which will be widely used by university admissions officers to evaluate the English proficiency of foreign applicants. In general, the more people who take the test and the more important the decisions which are based on test scores, the more care should be devoted to the construction of the test.

The following steps in test construction can be identified: (1) planning, (2) writing, (3) item selection, (4) test tryout, and (5) standardization. The first two steps, planning and writing the test, are common to the construction of all tests. The additional steps would be followed for the construction of relatively more "important" tests.

Planning The Test

Planning the test, which includes the specification of test content, is the single most important and in some ways the most difficult aspect of test construction. We have already identified the steps involved in specifying test content: (1) determining *what* it is we want to test and (2) determining *how* to test it. The second part of content specification, determining the types of task to be required, involves the determination of (1) the particular operations to be required of the examinee and (2) the structure or context embedded in these operations. The content of the test should be determined according to the purposes for which the test scores are to be used.

In addition to determining the content of the test, planning involves specifying the time available for the administration of the test and the number of items to be included. If we are assessing more than a single cube of the test specification framework, the number of items to be allocated to each cube should be specified. The number of items allocated to each cube should be determined in accordance with the importance or weight we wish to assign each cube. Thus, for example, if we are preparing a 100 -

item test covering four cubes, and if we wish to assign equal weight to each cube, then each cube should be represented by 25 items.

We should also specify the degree of difficulty which we want the test to have for the examinees, i.e., the percentage of examinees which we expect to answer each item correctly. The fewer the examinees who answer an item correctly, the greater the "difficulty" of the item. Conversely, the more who answer it correctly, the easier the item. Sometimes, we expect that most of the items contained by a test will be passed by most of the examinees. If we want to assess the degree to which our students have mastered material presented in class, for example, and if the material has in fact been mastered, then we can expect the test items to be very "easy," that is, most of the examinees will answer most of the items correctly. On the other hand, sometimes we want tests which are moderately difficult-tests whose items are passed, on the average, by approximately 50% to 60% of the examinees. We would want such tests if we were interested in identifying differences among examinees in their English proficiency, so that we could reliably say that one examinee's performance was better than that of a second or worse than that of a third. We would want to be able to rank the performances of our examinees if, for example, we were assigning grades. We cannot identify individual differences by means of test items which are passed by all the examinees or, conversely, by items which are failed by all of the examinees because if all the examinees perform in the same way, no differences can be observed. However, items which are passed by about half the examinees will maximize the identification of individual differences.

We have described two types of test in terms of item difficulty: one in which almost all of the examinees will pass each item and one in which about half will pass each item. Sometimes other distributions of item difficulty will be desired, depending on the purpose of the test. For a detailed discussion of the planning of item difficulty, the reader is referred to Thorndike (1949).

In summary, the steps involved in planning the test include the specification of (1) test content, (2) time for test administration, (3) number of items and their allocation to the different content areas, and (4) item difficulty.

Writing the Test

Once the test has been planned, items can be written to meet the specifications of the test plan. Item writing has been termed an art as well as a science. The novice item writer can avoid some of the more common pitfalls, however, by following the maxims for item writing offered by Thorndike and Hagen (1961). In addition, examples of foreign-language test items can be found in Lado (1961), Valette (1967), and Harris (1969). No matter how skilled the item writer, however, it is good practice to have the items reviewed by others. It is often difficult for an author to see faults in the items he has written, even though blemishes may appear obvious to a colleague. For tests which are to be standardized, it is a good rule of thumb to write twice as many items as are to appear in the final edition of the test.

Particular care should be expended on the preparation of the examinees' directions or instructions for taking the test. The preparation of unambiguous directions-instructions which will be uniformly and correctly interpreted-is a time-consuming but vital task. Care also should be taken in the preparation of directions to examiners.

Item Selection

By administering our items to a group of examinees who are representative of the people for whom the test is designed, it becomes possible to select, for a final edition of the test, items which have been found to be at the desired levels of difficulty. For example, if we have specified that our items should range in difficulty from 40% to 60%, with an average value of 50%, we can throw out all items passed by more than 60% of the examinees and all items passed by fewer than 40% of the examinees. We can then select from the remaining items those which will yield an average value of 50%. We can do this, of course, only if we have written enough "extra" items. If we find that we have not written enough items conforming to the desired statistical characteristics, we will have to choose those items which come closest to meeting our specifications in terms of percentage passing.

We can select each item not only on the basis of the percentage of examinees passing it but also in terms of the degree to which the item meets certain other criteria. For example, we would tend to reject an item which was passed by the poorer students (defined in terms of their score on the rest of the test or in terms of performance on some other criterion) but was failed by the better students.

We may also wish to select items in terms of the ease with which they are passed by native speakers of English who are otherwise similar to the examinees for whom the test is designed. If the item cannot be passed by native speakers of English who are otherwise similar to those for whom the test is written, it is likely that the item is badly written or is testing something in addition to, or instead of, English proficiency.

Test Tryout

Once the items have been selected for the test, a "dress rehearsal" of the final edition can be made with examinees who are similar to those for whom the test has been written and with examiners who are similar to those who will usually be administering the test. In the test tryout, we have a final opportunity to learn whether the directions to examiners and examinees are clear, whether the time limits are adequate, and whether the format of the test (the printing, arrangement of items on the page, etc.) and scoring arrangements are satisfactory.

Standardization

Standardization of the test enables users and prospective users of the test to determine the meaning and evaluate the usefulness of test scores. The steps involved in standardization include the collection of normative data and the gathering of evidence with respect to the reliability and validity of test scores.

Normative data enables the test user to interpret a test score in terms of the performance of one or more reference groups. A raw score-the actual number of items answered correctly-takes on meaning when we can compare it to the scores obtained by others. We would interpret a score differently if, for example, it represented the highest score obtained in a group of 100 students than if it represented the lowest. When we collect normative data, we attempt to describe the performance of various reference groups on the test so that such interpretations are possible. The reference groups we select should be chosen so as to be representative of the various groups who are likely to take the test, whether defined in terms of age, level of proficiency, first-language background, educational background, or type of prior language training.

The existence of data on reliability and validity enable us to judge the accuracy and the relevance of test scores. Reliability refers to the accuracy of stability of measurement. No language test can yield perfectly reliable scores. That is, examinees' scores (and their ranks in a group) would tend to change from one administration of the test to another if repeated measurements were possible. The data reported on reliability give us a means of estimating the stability of our scores-the extent to which an examinee's obtained (observed) score is close to his "true" score (the score which he would obtain on the average if it were possible to give him the same test repeatedly) or the extent to which the examinee would maintain his rank in the group if the test were given again. In general, the longer the test (the more items it contains), the more reliable it is. Thus the issue of reliability becomes more acute when we are dealing with short tests, particularly the relatively short subtests of a battery.

Validity refers to the relevance of the test scores to the purpose for which we plan to use the test. A test may be accurate without being relevant to our purposes. To take an extreme example, we may have a paper-and-pencil test containing 100 items requiring simple arithmetical operations, e.g., 4 + 3 =,4 - 2 =, 5 X 3 =, etc. Such a test may be quite reliable. It would also be valid if our purpose was to use it to assess the ability to perform simple arithemtical calculations. However, it would not be a valid measure of reading comprehension in English. Thus, validity must be judged in terms of the purpose of the test. Test publishers, as part of the standardization procedure, can provide data which will allow us to judge the test's relevance for our purpose. Such data are of varying types. A test of English language proficiency could offer data such as the correlation of test scores with other tests of English that are considered valid or correlation of test scores with other relevant variables, such as grades in English language classes, number of years in which English has been studied, number of years in an English speaking environment, or ratings by competent observers. Sometimes, test publishers offer, as evidence of validity, the degree to which the specification of test content and the correspondence of test items to the specifications are judged by authorities to be reasonable in view of the stated purposes for which the test was constructed.

Selecting a Test

In some cases, it may be possible to use a test which has already been developed. If a test has been developed for the same purpose and for the same or similar populations as ours, we can consider the possibilities of employing it. In part, we can judge the degree to which an existing test meets our purposes by evaluating its items. Do the items appear to tap the abilities that we wish to test? Do the items appear to be at the appropriate level of difficulty? Secondly, we can judge the test in terms of whatever data are available concerning reliability and validity. Third, we can judge the usefulness of the test in terms of the existence of norms. The test will be more useful to us if relevant normative data have been obtained. Finally, the usefulness of the prospective test can be judged in terms of such important practical considerations as the time required for administration, ease of scoring, cost per test, and the ease with which test security (preventing examinees from obtaining copies of the test) can be maintained.

Using the Test

No matter how good our tests, they can never be completely without error. All measurement involves some error, including the measurement of attributes such as second-language proficiency. Part of the difficulty in obtaining accurate measurement is due to the difficulty faced in adequately defining what is meant by second-language proficiency: hence, the importance placed upon specifying test content. Even with explicit specifications of content, however, language tests are by no means error free, and this fact should be constantly kept in mind when interpreting test scores or making decisions based on test scores. We should view test scores as hypotheses, not facts. We should view *differences* between test scores with even more wariness than we view a single test score, since the error contained by the difference between two scores is proportionately greater than the error contained by either of the scores alone. Thus, we should be quite cautious in hazarding the hypothesis that the difference between two scores is not due to chance. This is particularly true when we compare an individual's scores obtained on two subtests-for example between subtests of vocabulary and syntax-inasmuch as subtests are shorter and thus typically less reliable than scores based on the battery as a whole. Differences between foreign-language proficiency subtest scores should be viewed cautiously also because these scores tend to be related. In general, the greater the relationship between two scores, the greater the difference that must be observed between them before the difference can be considered real or not due to chance.

Whether we construct our own test or employ one which has been constructed by others, it is usually helpful to gather information about the test as we use it. Information can be collected about the individual test items-how many examinees pass each item and how well each item distinguishes the more proficient from the less proficient examinees. Such information can be useful when it comes time to revise the test or when other forms of the test are built. Information can also be gathered about the scores (on the test as a whole or on component subtests) of the various groups which take the test. Local norms-normative data based on the groups which we actually test-are ultimately the most useful ones and can be obtained by keeping records of the scores of those who take our tests. We should also keep records, if possible, of the degree to which decisions based on test scores have been found to be sound ones. For example, what proportion of the students admitted to a given program on the basis of test scores actually succeed in the program? Information about the way the test "works" can help us to (1) build better tests in the future, (2) interpret test scores more adequately, and (3) make better decisions based on test scores. We should not, in other words, unequivocally embrace a test once it has been constructed or selected. Tests can be improved, tests often should be improved, and the way we use them can be improved.

Summary

We can summarize the present chapter by restating the following points:

1. The use of foreign-language tests can be justified if decisions are made on the basis of test scores and if the information provided by the tests is relevant to such decisions.

2. The specification of test content is the single most important step in test construction.

3. Foreign-language test content should be specified in terms of (a) a three-dimensional framework comprised of the intersection of language variety, knowledge, and skill, and (b) specific tasks identified by the particular operations required and their structural or contextual representation.

4. The steps followed in the construction of a standardized second-language test include (a) planning, (b) writing, (c) item selection, (d) test tryout, and (e) standardization, which in turn involves the collection of normative data and data with respect to reliability and validity.

5. Selection (as contrasted to construction) of a test can be made in terms of (a) the degree to which the test user's objectives appear to be met by the test's items, (b) evidence on reliability and validity, and (c) certain practical considerations.

6. Test scores should be viewed tentatively, as hypotheses, not facts, about an individual.

7. Records should be kept with respect to the way the test functions in the setting for which it was selected or designed.

References

Buros, O.K. 1965. *The sixth mental measurements yearbook.* Highland Park, New Jersey: Gryphon Press.

Carroll, John B. 1961. Fundamental considerations in testing for English language proficiency of foreign students. In *Testing the English proficiency of foreign students.* Washington, D.C.: Center for Applied Linguistics. Pp. 30-40.

Carroll, John B. 1968. The psychology of language testing. In Alan Davies (Ed.). *Language testing symposium.* London: Oxford University Press. Pp. 46-69.

Chomsky, Noam. 1957. *Syntactic structures.* 's-Gravenhage: Mouton.

Chomsky, Noam. 1965. *Aspects of the theory of syntax.* Cambridge: M.I.T. Press.

Davies, Alan (Ed.). 1968. *Language testing symposium.* London: Oxford University Press.

Harris, David P. 1969. *Testing English as a second language.* New York: McGraw-Hill.

Pimsleur, Paul. 1966. Testing foreign language learning. In Albert Valdman (Ed.). *Trends in language teaching.* New York: McGraw-Hill. Pp. 175-214.

Thorndike, Robert L. 1949. *Personnel selection.* New York: Wiley.

Thorndike, Robert L. and Elizabeth Hagen, 1969. *Measurement and evaluation in psychology and education* (third edition). New York: Wiley.

Upshur, John A. and Julia Fata (Eds.). 1968. *Problems in foreign language testing. Language Learning,* special issue number 3, 1968.

Valette, Rebecca M. 1967. *Modern language testing: a handbook.* Harcourt.

WRITING PROFICIENCY AND ACHIEVEMENT TESTS

Sydney Sako

Teachers of English to speakers of other languages are often confronted with the problem of writing good test items both for the proficiency-type test and the achievement-type test. The technique of writing test items for the achievement test is similar to that for the proficiency test in that both are designed to measure the amount of skill in speaking and writing the language, and the level of listening and reading comprehension of the language.

Test items designed to measure the four different types of language skills as well as five different combinations or integration of these skills will be presented in this article. Various presentations of item types are also illustrated, such as the question, statement, sentence completion, short dialog forms, and their combinations. Also the best possible uses will be brought out of the different types.

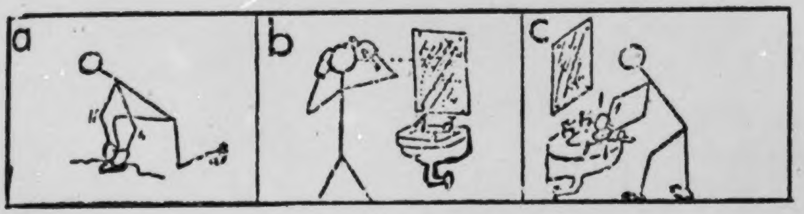
Test items can be written for each of the separate language skills as well as their combinations, such as:

1. Listening comprehension
2. Reading comprehension
3. Speaking production
4. Writing production
5. Listening-reading comprehension
6. Speaking production with listening comprehension
7. Writing production with listening comprehension
8. Speaking production with reading comprehension
9. Writing production with reading comprehension

1. Listening Comprehension

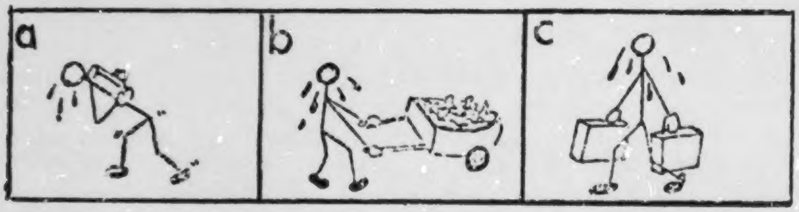
Items of listening-reading comprehension are often mistaken for items of pure listening comprehension. In order to test for straight listening comprehension there should be no reading involved; consequently, determination of a person's listening comprehension should be made through the use of pictures or movement of the body. The stem or lead can be presented in different forms, such as questions, statements, or their combinations. Examples of a picture test are shown as follows:

a. Question Form: Objective listening comprehension of *comb*.
Which person has a comb in his hand?

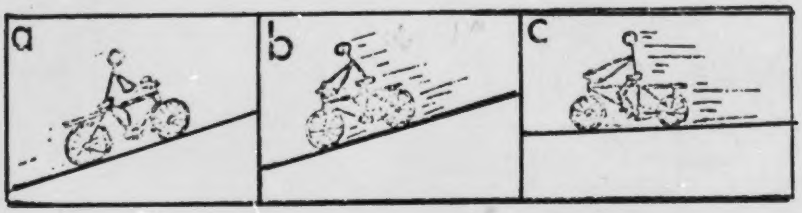


Sydney Sako is Chief, Measurement Branch, Defense Language Institute, Lackland Air Force Base, San Antonio, Texas 78236

b. Statement Form: Objective, listening comprehension of *on his back.* Joe is using both hands to carry the load on his back.



c. Statement-Question Combination: Objective, Listening comprehension of *level.*
Bill is riding on a level street. Which street is he using?



Some points to keep in mind when writing test items for listening comprehension are:

a. Does the stem present a clear, central problem?
b. Does the stem avoid giving away the correct answer by extra clues?
c. Is only the single objective being tested?

Examples of body motion test for listening comprehension:

1. Place your right hand on your forehead.
2. Wink your left eye.
3. Walk around the chair twice.
4. Open your book to page fiteeen.
5. Pick up a red pencil with your left hand.

The stems which are in command or request form are presented orally in an interview type situation. If the person responds appropriately, he is given credit for the item. The body motion test should contain at least 60 items to produce a comprehensive and reliable test. The response should be limited to actions that can be performed within the testing area.

The picture test and the body motion test measure the listening comprehension of an individual. These tests can be used for testing pupils and adults whose reading ability is limited; they can be used as sub-tests of a more comprehensive proficiency-type language test; or they can be used as an external criterion test for validation of other types of listening and reading comprehension tests.

The difficulty level of these items can be increased by adding action verbs of a more complicated nature, by adding idiomatic expressions, or by using symbols and objects which are more difficult for the examinee.

2. Reading Comprehension

The problems of constructing items for reading comprehension take on a different nature from those for listening comprehension. In the listening-type items the problems of pronunciation and controlled speed must be taken into consideration. Usually these items consist of conversational material. On the other hand, reading items usually represent samplings taken from the reading material found in textbooks, manuals, newspapers, magazines, technical books, and other sources. The vocabulary used, therefore, covers a wider range and the structure is more complex than that for the listening comprehension test. The stem may be presented in question form, statement form, incomplete sentence form, dialog, or in any of their combinations. Structure and vocabulary meanings are thus being tested. Examples of these items are as follows:

Lexical Items:

a. Question form:
What is a center of gravity?
  a. the midway point of any part of a body
  b. the point where the weights are equal
x c. the point about which all parts are balanced

b. Words in context.
The lever *oscillated* rhythmically.
  a. jumped in and out
x b. swung back and forth
  c. turned completely around

c. Statement form:
The enemy finally gave up.
x a. They surrendered.
  b. They attacked fiercely.
  c. They retreated.

d. Sentence completion:
He missed the question by giving the ______________ answer.
x a. wrong
  b. correct
  c. complete

e. Statement-question combination:
The survivors were located yesterday. Who were they?
  a. those missing
x b. those living
  c. those dead

Structural Items:

a. Sentence completion form:
Mr. Jones had thought that he ______________ next month.
x a. might arrive
b. had arrived
c. did arrive

b. Paragraph form:
New Orleans is located on the Mississippi River. At its numerous wharves, we can see the loading and unloading of large quantities of cotton, rice, sugar, and coffee. Because of its early history, it is one of the most distinctive cities in the USA. The oldest section is the French Quarter, where Bienville founded the city of New Orleans in 1718.

The city of New Orleans ______________
a. was originally called Bienville
b. is in the French Quarter
x c. is situated on a river

One of the important uses of the reading test is to ascertain the degree of reading comprehension of selected material by individuals regardless of their speaking ability or listening comprehension. A person at the elementary level will be able to understand only the basic structures and vocabulary limited to the basic needs, while a person at the professional competence level will be able to understand complex structures and vocabulary of a fairly sophisticated nature. If a person is required to pursue academic studies, instructor training, intelligence work, etc. he should have a high level of reading comprehension.

When writing test items for reading comprehension it has been proven valuable to consider the following questions:

a. Is the item, taken as a whole, specific?
b. Is the problem stated briefly?
c. Are the distracters important, plausible answers rather than obviously wrong?
d. Are all alternatives as well as the correct responses parallel in structure?

3. Speaking Production

A speaking test proposes to determine how well a person can speak the foreign language. In order to elicit oral responses, the directions or questions should be stated in simple terms. Stimuli to induce speech may take on the nature of a picture stimulus or perhaps the native language. For a speaking test, items are needed that will test the individual's pronunciation, fluency, vocabulary knowledge, and grammatical control.

Examples of types of items for a speaking test are as follows:

Objective	Type of Stimuli	Type of Response
a. Pronunciation	Written sentences	Reading sentences aloud
b. Retention and repetition	Sentences on tape	Repeating sentences aloud
c. Vocabulary and grammatical control	Picture	Description of the picture
d. Fluency	Interview or questions	Answer to questions

Specific segmental phonemes within each sentence should be evaluated for correctness in pronunciation, intonation, rhythm, and stress. The sentences on tape should be short and relatively simple at the beginning and longer and more complex toward the end. Students with low speaking proficiency may be able to remember and repeat the short sentences but not the long complicated ones. Students with high proficiency will be able to remember and repeat both. Simple sentences should be used in interviews so the student will not be penalized for failure to understand the directions or questions.

The student responses can be rated by the examiner by conducting individual interviews or by checking the recorded material. It is preferable to have a check list prepared in advance so that rating can be performed more objectively.

To rate the pronunciation the following rating scale may be used:

P-1 Mispronounces significant phonemes to such an extent as to be often incomprehensible.

P-2 Speech contains errors in some significant phonemes to the extent of requiring repetition, but is generally comprehensible. Accent very noticeable.

P-3 Student's accent is noticeable but not disturbing. His mispronunciation contains few irregularities in significant phonemes.

P-4 Student's pronunciation is adequate for immediate comprehension in normal conversation or situations.

P-5 Student pronounces clearly and distinctly and should be understood well over telephone as well as in face-to-face conversation. Has no peculiarities of accent or intonation which prevent immediate understanding.

To rate the fluency, vocabulary, and grammatical control, the following rating scale may be used:

E-1 Can formulate only simple phrases. Vocabulary and structure very limited.

E-2 Can formulate complete sentence structures within limited vocabulary. Can make himself understood in conversation of limited range.

E-3 Student's speech can be understood, but listener must pay closer attention than normal. Does not express himself easily, and must repeat or reword some statements to make them comprehensible. Sentence structure shows variations which hinder immediate comprehension.

E-4 Handles the standard speech patterns and structures with few significant variations.

E-5 Student's speech generally follows the normal English patterns and structures.

These ratings may be averaged, and then checked against the overall rating or against the rating done by another person. This will serve as a cross validation which tends to improve the reliability of the rating. Description of these ratings will help the test writer determine the difficulty level of the oral questions so he can write questions which would be appropriate for the desired level of the test.

4. Writing Production

In testing an indvidual's skill, elements such as vocabulary, structure, accuracy and speed of script writing, spelling, punctuation, content, and organization of material can be evaluated; however, the emphasis placed on testing the writing skills is usually on lexical and structural items, and on the accuracy and speed of script writing.

The stimuli used for eliciting writing production are similar to those prepared for eliciting oral production. They may be presented in picture form, in the native language, or in the target language in simplified form.

The grading of the written responses, like those of the oral responses, is also subjective. To improve the objectivity of scoring, graded samples of student's written material of the various levels can be used for comparison. The determination of the levels of these written materials may be made by individual teachers or by a panel of raters. A description of the different levels of the writing skills is presented below:

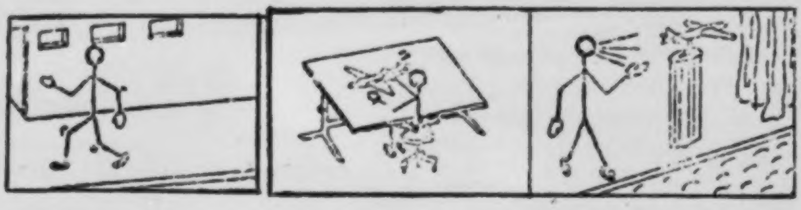
Description	Level
Elementary proficiency. Can write simple statements and questions using a very limited vocabulary. Errors in spelling and structure frequently obscure meaning.	1
Limited working proficiency. Can write sentences on familiar topics using non-technical vocabulary and basic structural patterns. Errors in spelling and structure occasionally obscure meaning.	2
Minimum professional proficiency. Can write paragraphs on familiar topics using basic structural patterns. Errors seldom obscure meaning.	3
Full professional proficiency. Can write prose with sufficient structural accuracy and sufficient vocabulary to satisfy professional requirements.	4
Native or bilingual proficiency. Can write with a proficiency equivalent to that of an educated native speaker.	5

The following is a sample listing of objectives, and types of stimuli and responses for the writing test:

Objective	Type of Stimuli	Type of Response
1. Accuracy and speed of script writing	Dictation (live or taped)	Writing dictated material
2. Vocabulary	Incomplete sentences	Writing the correct response which completes the meaning of the sentence correctly
3. Vocabulary and grammatical structure	Picture stimuli or instructions in the native language	Writing sentences or paragraphs

The following are sample pictures and simple directions for eliciting the writing responses:

a. Write what you see in the pictures shown below. Make complete sentences.



b. Write a paragraph on the topic "Why I want to study a foreign language."
c. Write an autiobiography of at least one full page.

5. Listening and Reading Comprehension

The listening and reading comprehension test is probably used more than any other combination to measure the integrated language skills. This type of test produces scores that correlate highly with success of predicting students' ability to acquire instruction in English. It lends itself well to the use of multiple-choice test items and therefore to the use of standard answer sheets which can be machine scored. The aural or listening portion is usually recorded on tape, while the reading portion is printed in the test booklet. The stem is presented orally in different forms, such as question, statement, incomplete sentence, or dialog form, or any combination thereof. The written responses provide stimuli for reading comprehension. The aural stems with the alternatives in written forms are illustrated as follows:

a. Statement form:
Mr. Jones took a bus to town (Aural)
x a. He rode the bus.
b. He put the bus away.
c. He pulled the bus to town.

b. Question form:
Did all the students do their homework last night? (Aural)
x a. Yes, they completed their homework.
b. Yes, they will do their homework.
c. Yes, they are working on their homework.

c. Dialogue form:
"Was Tom driving a car when you saw him yesterday?" (Aural)
"Yes, he was." (Aural)
x a. Tom drove a car yesterday.
b. Tom didn't drive a car yesterday.
c. Tom was riding in a taxi yesterday.

d. Incomplete sentence form:
The telephone is out of order. It is ____________ (Aural)
a. working
x b. not working
c. not installed

e. Statement-question combination:
The weather is variable. What kind of weather is it? (Aural)
a. cold weather
b. hot weather
x c. changeable weather

Care must be taken that items do not test memorization of facts, but mainly knowledge of the language:

Testing Memorization:	Testing Language:
The President is elected for __________	The President is elected for four years.
a. a. three years	a. He is appointed.
x b. four years	b. He is nominated.
c. five years	x c. He is chosen.

6. Speaking Production With Listening Comprehension

When a language teacher rates students by asking them questions and having them respond orally, he is testing listening comprehension and speaking production. The listening portion of the test should be designed so that very simple questions are asked at the lower level, and progressively more difficult questions at the higher level. Two methods can be used to determine the levels of difficulty of questions asked of the students:

1. Content determined by three variables: vocabulary, structure, and concept. The difficulty level of vocabulary can be obtained by frequency count, origin of word, and differences in meaning. The difficulty level of structure can be determined by the complexity of patterns used, by the senses, by the modes, and other grammatical usages. The difficulty level of concepts may range from the simple (concrete concepts) to the complex abstract concepts).

2. Item analysis using the ease index will reflect the percentage of the population answering the questions correctly. The questions comprising the lower level should have high ease indices, and those of the higher level should have low ease indices.

In other words, questions used in Level 1 should have very simple structure with basic vocabulary within the first 200 words and ease indices of 90 or over. Concepts used should be concrete in nature. Questions used in Level 5 should use complex structure, idiomatic expressions, and ease indices of .25 or less. They should consist of cultural items with abstract concepts, since this type of item requires greater knowledge of the language and culture, and a higher proficiency level.

Some sample questions which can be used in this test are listed under the five levels:

Level 1. Elementary Proficiency

1. Hello, how are you? (or Good Morning).
2. What is your name?
3. Where are you from?
4. Why did you come here?
5. What things do you see in this room?
6. Where do you eat lunch?
7. What time is it?
8. How do you go to town from here?
9. Do you drive a car?
10. Why do you go to the laundry?

Level 2. Limited Working Proficiency

1. Where did you go to school?
2. What are some of your work experiences?
3. What do you want to do in the future?
4. If you were very sick, what would you do?
5. How do you like to travel best?
6. What kind of climate do you have in your hometown?
7. What are some of the popular sports you like?
8. Do you plan to buy a car? What kind?
9. Who is your favorite movie star?
10. What is the meaning of the following paragraph?

(Read a paragraph and see if the student can explain it. Pick an easy one from a newspaper or magazine)

Level 3. Minimum Professional Proficiency

1. Describe the furniture you see in this room.
2. What are some activities you like to do best?
3. What is your specialty or profession?
4. Discuss your plans for the future.
5. Compare the weather here with that of your hometown.
6. What kind of tools or books do you use in your work?
7. How do you start your automobile?
8. What do you do when you put on the brakes?
9. What would you do if the lights would not work?
10. Listen to the following paragraph and explain the main points. (Choose a paragraph from his field of interest and read it in a normal way.)

Level 4. Full Professional Proficiency

1. Describe your educational background.
2. What do you like to do best when you are not busy?
3. Tell one of the most exciting experiences you have ever had.
4. Describe the person whom you admire most.
5. Tell of an emergency situation you have encountered in the past.
6. Have you ever run out of gas? What did you do?
7. Do you think there is a good chance of a hard freeze tonight?
8. Why is weather a common topic of conversation?
9. Have you ever broken in a pair of shoes?
10. Describe the kind of work you do?

Level 5. Bilingual or Native Proficiency

1. How long have you studied your second language?
2. What were your major courses in school?
3. Describe your work experiences in civilian life.
4. What do you like to do best on weekends?
5. What was one of the most pleasant moments of your life?
6. What do you mean by "home run" in a baseball game?
7. What would you do if you were told to cut in on Mr. Jones?
8. Did you ever back out of a bargain? Explain.
9. Did you ever get stuck on a purchase? Explain.
10. Why do we say that communication is a basic human activity?

The oral responses of different proficiency levels would also parallel the description just related. The same rating form used in the speaking production test will apply here. If a student answers the questions adequately, it is assumed that he understands the listening portion of the test. A student who answers the questions in an excellent manner will receive a plus to the rating (1-5). If he answers poorly, he will receive a minus to his rating. Rating students for listening comprehension and oral production is rather subjective, but a person can become more objective in rating through experience.

7. Writing Production With Listening Comprehension

The purpose of the test of writing production with listening comprehension is to check both the aural comprehension and the writing proficiency of the student. The main elements of writing that are usually tested are the proper choice of vocabulary and the correct usage of grammatical structure. Testing for the organization of content and the style of writing is usually made at the higher level of proficiency.

To check the aural comprehension and to elicit the written responses, the same types of questions mentioned in the aural-speaking test can be used, and these can be pre-recorded on tape.

Another type of oral presentation is the sentence completion form. This method of presentation requires the students to write the words or phrases which will complete the meaning of the sentence correctly. Examples of the sentence completion type are shown as follows:

1. A goat is a kind of ______________ (animal)
2. When water freezes, it turns into ________________ (ice)
3. A group of sheep is called a _______________ (flock)
4. A dull movie is one that is not __________________ (interesting)
5. If sugar is put in the coffee, the coffee is said to be ______(sweetened)

This type of test item requires greater skill in construction than others. There must be just enough information presented to elicit the required response. If not enough information is given, the examinee will not be able to come up with the intended correct answer. If too much information is given, the item will be a give away.

Not enough information:	Improved:
A sheep is _____________.	A sheep is a kind of ______________

If an item can elicit several responses which can be correct semantically and syntactically, credit should be given to the student when he responds to any of the possible answers.

8. Writing Production with Reading Comprehension

Some of the quizzes used in the classrooms or portions of the proficiency tests may be classified as tests of writing production with reading comprehension. The purpose of this type of test is to determine how well a person can understand the reading material and at the same time to determine how well he can write the answers. The same type of topics or directions used in the reading-speaking test may be used for this test. If a student has had the same amount of training in oral as well as in writing skills, he should do just about as well in one test as the other. However, if the writing symbols of the student's native language are drastically different from those of the language he is learning, such as a Chinese speaker who is learning English, but is accustomed to writing ideographs that show no phonemic contrasts, then such students may do better in the speaking than in the writing test.

Two types of criteria may be used to determine the levels of reading and writing proficiency. One criterion is to use the rating scales, and the other is to compare the student's written material against a graded sample. The grading of this type is subjective. We can make this test semi-objective by using incomplete sentences. We find more freedom in the placement of blanks in the incomplete sentence form in this test than in those used in the listening-speaking test.

The following are some examples of the completion type sentences that may be used in the reading-writing test:

1. Mary read the letter carefully, didn't ____________________? (she)
2. He is ______________most likeable person I know. (the)
3. What time _________________he get up yesterday? (did)
4. Do you mind _______________I smoke here? (if)
5. It's very stuffy in here; please turn ___________the fan. (on)

Some other examples of test items which can be used in this type of test are as follows:

Change the statements to questions. Keep the same tense.

Statements	Questions
1. This is your book. _____________	(Is this your book?)
2. He was not in class today. _____________	(Wasn't he in class today?)

Change the sentences to negative sentences.

Sentences	Negatives
1. This is my pen. ________________	(This is not my pen.)
2. He saw an accident yesterday. ________________	(He didn't see an accident yesterday.)

9. Speaking Production With Reading Comprehension

When a written topic is presented to a student and he is required to tell something about it, or when he has to read a passage and then say something about it, we have a speaking test with reading comprehension. The student must be able to understand the written portion and at the same time be able to respond orally. The same questions used in the speaking test with listening comprehension may be used here. Of course, the students will read the questions instead of listening to them. Examples of other topics or directions which can be presented to the students are as follows:

a. My favorite food.
b. My most embarrassing moment.
c. My earliest recollections.
d. Describe one of the largest cities you have lived in. What things interested you the most?
e. Describe a method of transportation you like best. Why do you pick this method.
f. Tell something about your favorite sports. Why do you like them?
g. Talk about some interesting things concerning your home town, including the location, historical events, people, culture, educational system and industries.

h. Talk about your interests in life, such things as your family, educational experiences, work experiences, hobbies, and future plans.
i. Talk about the climate of the place you came from, temperature range, rainfall, extreme fluctuations of the weather.
j. Talk about some interesting people you know, including your friends, favorite movie stars, former teachers, and some outstanding scientists or political figures.

Conclusion

The use of the different types of items for testing proficiency and achievement presented here will depend upon the objectives of the training course. If the student is expected to use his speaking ability rather than his writing ability, the speaking production test or the speaking production test with listening comprehension will be used almost exclusively. If the student is expected mainly to understand the lectures and reading material, it is recommended that a listening-reading comprehension test be used. Thus the selection of the types of items and their use will ultimately be based on the objectives of training and the needs of the student; and teachers of English to speakers of other languages will play a vital role in checking these objectives and needs, and finally in constructing the proper instrument for evaluation.

The Assessment of Language Dominance and Proficiency[1]

Marina Burt and Heidi Dulay

INTRODUCTION: DIMENSIONS OF BILINGUAL MEASUREMENT

In this paper, bilingual child refers to any child who lives in a home where a first language other than English is spoken by one or more members of the household.[2] A child's degree of bilingualism, then, may range from a point approximating monolingualism -- where the child may only comprehend the most basic syntax and vocabulary in one language while speaking another fluently -- to a state of "balanced" bilingualism where the child's proficiency in both languages is equally high.

The identification and diagnosis of bilingual children's linguistic backgrounds, performance, and environment are tasks that face countless school personnel across the country who are responding to requirements of the Lau v. Nichols decision of the U.S. Supreme Court (414 U.S. 563, 1974), federal statutes such as Title VII of the Elementary and Secondary Education Act, and state legislation such as California's AB 1329. Students who are the designated recipients of mandated special educational services are described by various terms, such as "non-English speaking" (NES), "limited English speaking" (LES), "non-English dominant minority," "linguistic minority," etc. Although the appropriateness of this terminology is subject

---

Heidi Dulay works at Bloomsbury West, Oakland, California 94618
Marina Burt is President of Bloomsbury West, Oakland, California 94618

[1] This paper is a revised version of an earlier paper: Burt, M., H. Dulay, and E. Hernandez Ch., "Evaluation of linguistic proficiency in bilingual children," in Singh, S. & J. Lynch (eds.), Diagnostic Procedures in Hearing, Language and Speech (Baltimore, Md.: University Park Press) in press. Sections of this paper have been reprinted with permission of University Park Press.

[2] According to the Survey of Languages Supplement to the July 1975 Current Population Survey, more than 7.6 million school age children in the U.S. live in households where languages other than English are spoken, in addition to the 17.7 million who are 19 years old or older living in such households (Waggoner, 1976).

to continuing controversy, and consensus on the definitions of these terms is difficult if not impossible to attain, implicit in the definitions of all of the terms are one or more of the following dimensions of bilingual measurement:

1. Language dominance assessment
2. English proficiency assessment
3. Other language proficiency assessment
4. Home language identification
5. Primary language identification

These dimensions are useful not only for reporting purposes, but also for effective educational planning and program evaluation. We briefly discuss them here in order to provide a context for the narrower focus of this paper namely, the assessment of oral language dominance and proficiency.

Language proficiency refers to the degree to which an individual exhibits control over the use of the rules of a language for one, some, or all of its numerous and diverse aspects: phonological, syntactic, and lexical systems, orthography, discourse and stylistic rules for oral and written communication in various social circumstances. One may evaluate students' English proficiency in any one or more of these aspects, or their Spanish proficiency in any other language for any of these aspects. The results of such measurement would yield proficiency information for the students in the particular language tested. Proficiency in one language, of course, does not necessarily imply lack of proficiency in another, and conversely, lack of proficiency in one does not imply proficiency in another.

Language dominance refers to an individual's degree of bilingualism, that is, the relative proficiency of an individual with respect to more than one language. Thus, one might measure a student's proficiency in two languages independently, then compare the results to obtain that student's language dominance. A child would be, say, "Spanish dominant" if the child were more proficient in Spanish than in English. Or, a child would be "balanced" if the level of proficiency attained in both languages were the same.

Dominance in one aspect of languages does not imply dominance in the others. For example, a child's command of the phonological system of the first language may be greater than that of the second, but the child may have a larger active school vocabulary in the second language than in the first. In measuring dominance, skills measured in one language must also be measured in the other. For example, if the oral production of syntactic structures is measured in one language, then the oral mode must also be used to measure syntactic proficiency in the other language. Oral skills are the most commonly used as the basis for measuring dominance. Other skills, such as reading, are usually measured independently (Cornejo et al., 1976) since reporting requirements, instructional planning and curricular activities for oral language development and reading differ considerably.

Another dimension of bilingual measurement is the identification of a student's home language(s). This involves determining which languages are used in a student's home environment, how much they are used, and sometimes, who uses which languages. The use of a given language in the home by parents or others residing there does not necessarily imply that the student is proficient or dominant in that language. The precise relationship between the use of a language in the home and student proficiency in that language is still a research question.[3] For example, in a recent study, Ramirez and Politzer (1975) found that while the use of Spanish in the home correlated well with children's Spanish proficiency, the use of English did not. Many times, however, despite parents' or grandparents' use of non-English language at home, and despite their special efforts to speak with their children in the "home language," children respond in English, and, over a period of time, lose all but an ability to comprehend the non-English home language. Thus the identification of home language use serves to supplement (not replace) information on a student's language dominance and proficiency. Information on linguistic characteristics of the home is important input for sound instructional planning.

A student's primary language is simply the language the student acquired first, usually during the first four or five years of life. (In the research literature the term "first language" is more common; however, in legal documents, "primary language" is preferred.) In some cases, two languages may be acquired simultaneously during early childhood, in which case the child may be said to have two primary languages.[4]

All of these dimensions of bilingual measurement play specific roles in educational planning and evaluation, and in responding to reporting requirements for local, state, and federal government agencies. Table 1 summarizes some of their most common applications.

---

[3] D. Waggoner at the National Council of Educational Statistics is currently conducting a research study on this topic, in cooperation with the Educational Equity Group of the National Institute of Education.

[4] The dynamics of simultaneous bilingual acquisition has been the subject of research studies in recent years (Padilla & Lindholm, 1976; Swain, 1972). The more commonly reported phenomenon, however, is the sequential acquisition of two languages, where the process of second language acquisition begins after the basics of the first (or primary) language have been acquired.

TABLE 1
Applications of Bilingual Measurement

LINGUISTIC DIMENSION MEASURED	APPLICATION
Language Dominance	1. Bilingual program planning, including student and faculty assignment. 2. Placement in non-English or English medium reading and subject matter classes. 3. Initial diagnosis to determine language in which further testing is to be conducted. 4. Program evaluation and needs assessment for funding. 5. Reporting to the Office for Civil Rights, the State Department of Education and other agencies.
English Proficiency	1. Placement in appropriate English instructional group. 2. Determination of readiness to begin English reading instruction. 3. Program evaluation and needs assessment for funding.
Other Language Proficiency	1. Placement in appropriate Other Language instructional group. 2. Determination of readiness to begin reading instruction in the other language. 3. Program evaluation and needs assessment.
Home Language Usage	1. Educational planning, including choice of bilingual program model and determination of parental participation and assistance. 2. Determination of need to notify parents of school activities in non-English languages. 3. Reporting to the Office for Civil Rights, the State Department of Education, and other agencies.
Primary Language	1. Reporting to the Office for Civil Rights, the State Department of Education and other agencies.

A full discussion of each dimension is far beyond the scope of this paper. We have chosen, therefore, to focus on the assessment of oral language dominance and proficiency --- the critical first steps in the placement of children into appropriate reading and subject matter classes in bilingual programs. Three subtopics will be discussed: global guidelines in evaluating language dominance and proficiency instruments, selection of the linguistic aspect(s) to be measured, and types of oral language elicitation tasks.

SOME GENERAL GUIDELINES FOR EVALUATING ORAL LANGUAGE PROFICIENCY AND DOMINANCE TESTS

The psycho- and sociolinguistic aspects of becoming bilingual have just begun to attract fairly widespread attention. Because of the recency of this interest there has not been sufficient time to develop many instruments which incorporate the research findings concerning the development of the second language, the first language, and the interaction of the two described in recent literature. (See Dulay, Hernandez & Burt, in press; and the articles in Burt, Dulay & Finocchiaro, 1977 for current literature reviews.) We therefore offer some general criteria for evaluating oral proficiency tests or procedures used to measure language dominance. These criteria reflect both the recent research and our current state of knowledge concerning oral language testing for bilingual children. The guidelines are not intended to be exhaustive, but rather to serve as partial criteria for evaluating oral proficiency and dominance instruments.

1. The parts of a language dominance test that assess each language must not be translations of each other.

When one compares the grammar of any given two languages, numerous differences may be observed in various surface aspects of the languages. For example, the morphology of Spanish includes gender markings on articles and adjectives: *la manzana roja* requires the feminine article *la* (as opposed to the masculine *el*) and the feminine marker on the adjective *roja* (as opposed to the masculine *rojo*), since the head noun is feminine. One of the rules to be acquired by learners of Spanish therefore, is gender agreement on articles and adjectives when these modify nouns. In the case of English, no such gender agreement for articles and adjectives exists. Thus, the question or item used to elicit a response such as *la manzana roja* that is designed to see whether a child has acquired gender agreement in Spanish would not be useful in the evaluation of English proficiency. (In *the red apple*, *the* and *red* do not change their form when the head noun apple changes to some other noun, e.g, flower.) Alternatively, one may wish to test a child's control of the *do/did* distinction in certain English constructions (e.g.,*Yesterday she do/did not go out*). Since the *do* auxiliary does not exist in other languages, the English question used to elicit it would not yield useful information about another language were it translated.

These examples, and many others like them, demonstrate why the different language versions of a dominance test must not be simple translations of each other. The linguistic structures of languages differ, and structural distinctions made in one language may not be the same as those in another. The corresponding language sections used in a dominance test must therefore be based specifically on the syntax (or phonology, etc.) of the particular language being tested. In some cases, of course, there will be similarities across languages for some structures, e.g., plural formation for nouns in

English and Spanish (*window-s*, *ventana-s*). In those cases similar questions may be used to elicit analogous structures. Such similarities must be determined *prior* to construction of the test items in each language, however, so that items accurately reflect the structures and their development in a given language.

2. *The content of a language measure must not be outside the student's experience or cultural customs and values*.

In order not to confound linguistic proficiency and knowledge of the world, the content of a language measure (the concepts and activities depicted in pictures or referred to in questions) must not be outside the experience of the students being tested, nor inconsistent with their cultural customs and values. For example, a northern winter scene might include pictures of skis, sleds, a snowman and a snowball fight. If such pictures are included as stimuli, students who have never experienced a northern winter would be penalized unfairly when they are unable to answer questions based on these unfamiliar things. Likewise, a green banana represents an unripe banana to some people, but to others it represents a variety of banana used for cooking. Thus, if the target "descriptors" to be scored for green and yellow bananas include *unripe* and *ripe*, some students would be unfairly penalized: their failure to include the desired descriptors would *not* be due to a language deficiency but to a perception of bananas that is different from the test constructor's. Likewise, students who are unfamiliar with penguins will not be able to answer correctly questions such as "Do penguins waddle?" Responses to test items that rely on unfamiliar content or that assume only one of several possible interpretations do *not* reflect general language ability, but merely indicate the child's exposure to that particular content.

3. The responses required by test items must not violate conventions of natural discourse.

If a complete sentence is required (for a perfect item score), then questions to which a single word is a natural response must not be used. For example, a question such as "Who is this?" may be, and often is, answered with just a name or a simple noun phrase. Similarly, questions such as "Is this a pencil?" are *typically* answered by native speakers of English with a simple "yes" or "no" rather than with "Yes, this is a pencil" or "Yes, it is." Yet, many language tests require students to produce complete sentences in response to such items in order for the response to be scored as correct or to be scorable at all. Clearly, such requirements result in eliciting *artificial* rather than natural speech. It is unfair to penalize students for their failure to produce complete sentences when the test questions fail to elicit them naturally. Language elicitation tasks are discussed more fully below.

4. A distinction must be made between the quantity and the quality of the student's responses.

The use of open-ended questions such as "Tell me all you can about this picture" tends to obscure the critical distinction between quantity and quality, since some students respond enthusiastically to such a question, spinning endless tales, while others give limited responses in order to avoid saying something wrong (Cazden, 1972, Labov et al., 1968). When students are not familiar or comfortable with the examiner, or if they are not certain of what is being asked, they will tend to give short and limited responses (often followed by a nervous "right?"). Open-ended and imprecise questions lead to rewards for superficial verbosity and to penalties for those students who are less talkative or shy (especially with strangers), or who do not understand the intent of the question. Ironically, students who are more advanced linguistically may be penalized for short but appropriate responses in these situations. "Tell me all you can" questions may also confuse language proficiency with the student's comprehension of the task, attention to detail in a picture, ability to create a story, or volubility, unless, of course, these factors do not affect the scoring of a student's response.

5. Age and grade norms cannot be used in interpreting bilingual test scores.

The interpretive framework of language tests for bilingual students cannot follow the norm-referenced approach of interpreting student performance in relation to specific age or grade reference groups, unless length and type of exposure to the target lanquage can be incorporated into the norms. That is, neither the kind of English nor length of exposure to English is predictable by the age or grade of the student. For example, a second grader in the U.S. who has hardly been exposed to English cannot be expected to have reached the average level of English proficiency attained by other second graders who have been exposed to English normally all their lives. Likewise, second graders whose bilingual parents choose to speak English at home cannot be expected to maintain Spanish proficiency to the same degree as children whose parents speak to them in Spanish most of the time.

The inclusion of language exposure variables into age and grade norms would be extremely complicated, if not impossible. At present, norms including these crucial variables do not exist. The research required is extremely complex and has only recently begun. Type of exposure to a target language includes variables such as the use of the language at home by parents and siblings, interaction with peers who speak the target language fluently, use of the language in the neighborhood, and even the amount of television viewing. These factors vary from community to community as well as from student to student, making it extremely difficult to obtain meaningful norms in the traditional sense.

In view of these complex problems in establishing age and grade norms for the language proficiency of bilingual students, it seems that non-normal language functioning is most clearly indicated by a student's poor performance (low scores) on tests of proficiency in both first and second languages. For example, on the Bilingual Syntax Measure (BSM), a student's proficiency level may range from a low of Level 1 to a high of Level 5 in either the first or second language. Someone who scores in Level 3 ("survival level") or lower in both languages assessed is classified in the "Special Diagnosis" category, rather than in any one of the normal bilingual categories. (In most of the normal categories of bilingual development, proficiency in one language is at least one level lower than the other, "balanced bilingualism" being the only exception.)

In addition to the above discussed considerations, much rethinking has gone into the issue of what among the various aspects or parts of language should be - and can validly be - tested. The following section summarizes those considerations.

SELECTION OF THE ASPECT(S) OF LANGUAGE TO BE TESTED

The four aspects of language which have been most commonly assessed by language proficiency and dominance instruments, or which are usually included in second language curricula are: vocabulary, pronunciation, syntax (including morphology), and functional use. Advantages and disadvantages of each as indicators of general language development in bilinguals will be discussed in the following sections.

1. Vocabulary

One of the most traditional ways of measuring verbal ability is the vocabulary test. The assumption is made, either explicitly or implicitly, that students for whom vocabulary tests are appropriate will have had those experiences which provide occasions for learning certain vocabulary items, certain thinking skills, and other linguistic skills which correlate highly with range of vocabulary. Vocabulary tests are thus most appropriate for testing groups which have had homogeneous experiences.

But contemporary educational needs call for a test which is appropriate for children from heterogeneous cultural, social, and linguistic backgrounds, and whose languages have been learned in diverse environments. Since word knowledge is a function of experience, there is no assured body of vocabulary to which all these children will have been exposed. Regional vocabulary differences, for example, often include just the common, familiar lexicon that appears in elementary tests of language proficiency; and more often than not, it is just these words that are likely to differ from region to region and between the standard and the local variety. For example, in Spanish the words for hat, jacket, banana, beans, peas, cooking, shaving, driving, and hundreds more will find different expression depending upon the locale and the mode of use. For a heterogeneous population, therefore, the use of vocabulary as an indicator of language development penalizes some students for not having been exposed to the "right" vocabulary.

Nevertheless, the growth of a child's vocabulary is indeed a necessary part of linguistic development, and consequently, an important part of a language curriculum. In order to fairly and accurately assess a bilingual child's lexical progress, criterion-referenced instruments developed for particular groups of children would be the most appropriate assessment tool. Such tests would be school achievement tests rather than tests of intelligence or general linguistic development.

2. Pronunciation

Although pronunciation is a commonly measured aspect of language, it has serious drawbacks if it is used as a major indicator of normal bilingual development. There is a great deal of variability in pronunciation across idiolects and dialects, not only in the second language, but in the first as well. For example, in Spanish there are numerous variant pronunciations according to country and region.[5] Not only do many of the individual sounds differ, there are important differences in intonational phenomena. Particularly important are pitch and rate of speech differences, as well as certain junctural phenomena such as (?) (glottal stop) or velarization of /n/. (We might compare these kinds of pronunciation differences to those between Midwestern American speakers and native English speakers from, say, New Delhi or Aukland.) Furthermore, it is not uncommon that a bilingual maintains a "foreign accent" in the second language, even when native-like syntactic and lexical proficiency have been achieved.

Extreme care must be taken therefore, not to confound differences in pronunciation with difficulties requiring remediation. Accent is more often an indicator of regional background, socioeconomic status, and ethnic or peer group affiliation than of general language proficiency. In particular, bilingual speech communities have norms of pronunciation for both languages that may differ from those used by monolingual speakers of either language.

---

[5] We refer here to the standard language only. Additional differences exist when variation in social dialects is considered. (See Hernandez-Chavez, 1975, which includes a collection of important articles on Chicano Spanish.)

3. Functional Use

Another aspect of language consists of its functional uses -- asking questions, giving commands, making promises or apologies, addressing superiors, etc. Although these forms are a basic part of communication, they are difficult to elicit systematically, efficiently, and naturally from large numbers of students. It is almost impossible, for example, to elicit a sincere apology without saying something artificial such as "please apologize to (somebody) for (something)." Creating situations in which almost any student would feel naturally compelled to apologize in a given way (in order for the test to measure the same speech act) is also most difficult, if not impossible. Because of differences in environment and social custom, it is not presently possible to construct the number of natural situations that would be required to test knowledge of such speech acts or discourse rules. Complicating the matter further is the fact that there are many acceptable ways to execute a speech act, depending upon the speaker's culture and various features of the sociolinguistic context.

Assessment difficulties caused by intergroup variability are especially prominent in the case of a multicultural population spanning a wide socioeconomic range, such as that with which we are concerned. As was mentioned for vocabulary, however, learning the functional uses of linguistic structures comprises an important component of language development. This aspect of linguistic development might be assessed more fairly and accurately by criterion-referenced instruments designed for a particular locale or curriculum.

4. Syntax

This aspect of language is usually referred to as the structural backbone of any language. In broad terms, syntax may be defined as the system of rules for the arrangement, interrelationship and form of words, phrases, and sentences in a language. Of the various aspects of a particular language, syntax shows the least variation among speakers of the language. For example, one does not typically find that the position of morphological endings or constituents changes in different geographical regions. For example, The dog ate it does not change to Dog the ate it or any other possible combination of those words anywhere in the U.S. For most of syntax, we are assured that students have not had variable exposure. Where variation does exist, it is usually systematic and can therefore be incorporated into the interpretive system of language instruments.

Most research in child second language acquisition has focused on syntax. As a result, descriptions of syntactic developmental sequences that obtain across groups of students from diverse language backgrounds and learning environments are available. (See Dulay & Burt, 1975; Krashen, 1975 and 1977; Fathman, 1975; and Burt, Dulay & Hernandez, 1976.) These, along with information on common error types, provide invaluable input for the assessment of language development in bilinguals.[6]

For these reasons, syntax seems to be the most suitable and least biased aspect of a language to rely upon in the evaluation of general linguistic proficiency of bilingual multicultural students. Of course, in relying primarily on syntax, other aspects of language are excluded. But given the present state of the art, this may be unavoidable when one is primarily interested in the fairest and most reliable indicator of general language development in normal bilinguals.

---

[6] The Bilingual Syntax Measure is the only test to date which incorporates syntactic and morphological developmental sequences into its level of proficiency.

ORAL LANGUAGE ELICITATION TASKS

Of the four language skills -- comprehension (listening), production (speaking), reading, and writing -- production is the most commonly used mode for determining the linguistic dominance and proficiency of bilingual students. In this section, therefore, we focus on the types of tasks used to elicit oral language for purposes of assessment.

Most tasks used to elicit speech samples or verbal responses may be grouped according to the presence or absence of a communicative focus. Accordingly, we label them natural communication tasks and linguistic manipulation tasks, respectively. A natural communication task is one where the focus of the student is on communicating something to someone else -- an idea, some information, or an opinion in a natural manner. In such situations the speaker unconsciously uses the grammar rules acquired to convey the message. For example, a question such as Why do you think he's so fat? (asked by an examiner while pointing to a fat character) elicits an opinion or idea from the student which is directed towards that specific situation. The resulting speech is produced by the student with little, if any, conscious focus on its linguistic form.[7] The speech is offered in order to communicate the student's opinion to the examiner.

---

[7] See Krashen, 1977 for a discussion of conscious "monitoring" in second language performance.

On the other hand, a linguistic manipulation task is one where the focus of the student is on performing the conscious linguistic manipulation required by the task. For example, asking a speaker to transform No one was here into a yes/no question requires manipulation of the elements in the sentence; however, the activity in itself does not serve any communicative function for the student. Rather, the student is consciously focusing on the linguistic rules required to perform the operation requested, an activity which is rarely a part of natural communication.

In sum, natural communication tasks permit one to make statements concerning the student's normally developing (and unconscious) grammar, while linguistic manipulation tasks permit one to make statements concerning the student's meta-linguistic awareness, i.e., the conscious knowledge of and manipulation of the rules and forms of a language.

Below, we present common examples of both types of tasks. Natural communication tasks include structured and non-structured communication (see Tables 2 and 3), while linguistic manipulation tasks include imitation, translation, substitution, completion, and others.

Table 2 provides a comparison of natural communication and linguistic manipulation tasks in terms of their definitions, advantages and disadvantages with regard to their appropriateness and efficiency as indicators of linguistic proficiency.

TABLE 2

Comparison of Two Major Types of Oral Language Elicitation Tasks: Natural Communication and Linguistic Manipulation

	NATURAL COMMUNICATION	LINGUISTIC MANIPULATION
DEFINITION	1. Taps student's unconscious use of grammatical rules to produce utterances in a conversation. 2. Uses natural speech where student's focus is on communicating something.	1. Taps student's conscious application of linguistic rules to perform a non-communicative task. 2. Uses artificial "speech" where student's focus is on a given rule.
SOME TYPES	Structured communication, Non-structured communication etc. (See Table 3)	Imitation, Translation, Completion, Transformation, Substitution, etc.
ADVANTAGES	1. The language sample obtained represents natural communication, the skill that is ultimately being assessed. 2. The task is virtually free of confounding task biases.	1. Target structures seem to be readily obtained.
DISADVANTAGES	Certain structures are extremely difficult to elicit naturally, e.g.,perfect tenses (had seen).	Confounds conscious knowledge and use of grammar rules with ability to use the language for communication.

1. Natural Communication Tasks

Table 3 compares two major types of natural communication tasks: structured and non-structured. As the table is self-explanatory, we will not discuss it further.

TABLE 3

Comparison of Structured and Non-Structured Natural Communication Tasks

	STRUCTURED COMMUNICATION	*NON-STRUCTURED COMMUNICATION*
DEFINITION	Natural conversation between s student and examiner where examiner asks student specific questions designed to elicit target structures naturally and systematically.	Natural conversation between student and examiner or other person where there is no intent to elicit specific structures.
ADVANTAGES	Target structures may be elicited selectively and quickly; more efficient than non-structured communication.	Structures that are difficult to elicit with specific questions may be offered by subjects spontaneously.
DISADVANTAGES	Not all structures are easily elicited, e.g., yes-no questions.	1. A great deal of speech must usually be collected before a sufficient range of structures is used by the student to permit assessment of linguistic proficiency. 2. One cannot make any statements about the student's control over structures not offered during the collection periods (since one cannot be certain why a structure was not offered, i.e., whether the situations did not require it or whether the child did not know it)

2. Linguistic Manipulation Tasks

Generally speaking, (as mentioned in Table 2) all non-communication tasks share at least one major deficiency when used as indicators of general linguistic proficiency, namely, they require student behaviors which are not typical of natural language use. They are therefore severely limited as general proficiency or dominance indicators. To clarify this, consider some aspects of two commonly used linguistic manipulation tasks: imitation and translation.

Imitation: the student repeats an item or sequence of words, sounds, phrases or sentences after a model. Imitation tasks may reveal some structures the student does not yet control and may provide information about the student's short term auditory memory. On the other hand, imitation tasks do not permit inferences about which structures the student controls, since students may correctly imitate structures and forms that they do not yet control in spontaneous communication. (For example, a child may repeat something correctly in class but use it incorrectly on the playground.) Imitation may also confound memory or attention span with linguistic proficiency. For example, a proficient student may err in repeating a sequence of words (or sounds) just heard because of a short memory or attention span rather than low proficiency. Likewise, students with better memories and longer attention spans, but who are not yet proficient, may be able to imitate a sequence of words accurately, obscuring an actually lower level of proficiency.

Translation: the student provides another language equivalent of an utterance provided. Translation tasks may encourage word-for-word matching between the student's first and second language. The errors that may result would appear to be due to interference from the first language when in fact they may have been induced by the task. This problem is especially obvious in timed translation tasks, where the first step for many is word matching. When time does not permit the student to reorganize the translation to make it "sound natural," an erroneous impression of the student's proficiency in the target language results.

As interest in the measurement of linguistic proficiency and dominance grows, we may expect there to be a commensurate amount of research directed towards discovering more ways to measure a student's ability to communicate naturally in diverse situations. We believe that present efforts directed towards discovering ways in which to assess the acquisition of natural communicative skills have already and will continue to cause fundamental changes in the construction of instruments to measure language dominance and proficiency.

REFERENCES

Burt, M. & H. Dulay (eds.). *New Directions in Second Language Learning, Teaching, and Bilingual Education*. Washington, D.C.: TESOL, 1975.

Burt, M., H. Dulay & M. Finocchiaro. *Viewpoints on English as a Second Language*. New York: Regents Publishing Company, 1977.

Burt, M. H. Dulay & E. Hernandez Ch. *Bilingual Syntax Measure* (English, Spanish, and Philipino Editions). New York: Harcourt Brace

Burt, M., H. Dulay & E. Hernandez Ch. *Technical Handbook: Bilingual Syntax Measure*. New York: Harcourt Brace Jovanovich, 1976.

Cazden, C. *Child Language and Education*. New York: Holt Rinehart & Winston, 1972.

Cornejo, R., L. Trithart & D. Gonzales. Prediction of reading in English for bilingual students. Paper submitted to AERA, San Francisco. Bilingual Evaluation Systems, Center for the Study of Evaluation, Graduate School of Education, UCLA, 1976.

Dato, D. (ed.). *GURT 1975: Developmental Psycholinguistics Theory and Applications.* Washington, D.C.: Georgetown University Press, 1975.

Dulay, H. & M. Burt. A new approach to discovering universals of child second language acquisition. In Dato, D. (ed.), 1975, 209-233.

Dulay, H. & M. Burt. Remarks on creativity in language acquisition. In M. Burt, H. Dulay & M. Finocchiaro (eds.), 1977, 95-126.

Dulay, H., E. Hernandez Ch. & M. Burt. The process of becoming bilingual. A chapter in S. Singh & J. Lynch (eds.), in press.

Fathman, A. Language background, age and the order of acquisition of English structures. In M. Burt & H. Dulay (eds.), 1975, 33-43.

Hernandez Ch., E., A. D. Cohen, & A. Beltramo (eds.). *El Lenguaje de los Chicanos: Regional and Social Characteristics of Language Used by Mexican Americans*. Arlington, VA: Center for Applied Linguistics, 1975.

Keller, G.D., R.V. Teschner, & S. Viera (eds.). *Bilingualism in the Bicentennial and Beyond*. New York: Bilingual Press, 1976.

Krashen, S.D., C. Madden & N. Bailey. Theoretical aspects of grammatical sequencing. In M. Burt & H. Dulay (eds.), 1975, 44-54.

Labov, W., P. Cohen, C. Robbins & J. Lewis. *A Study of the Non-Standard English of Negro and Puerto Rican Speakers in New York City*. Final Report of Cooperative Research Project No. 3288, Columbia University. Available through ERIC. 1968.

Nakano, P.J. Educational implications of the *Lau* v. *Nichols* decision. In M. Burt, H. Dulay & M. Finocchiaro (eds.), 1977, 219-234.

Padilla, A.M. & K. Lindholm. Acquisition of bilingualism: An analysis of the linguistic structures of Spanish/English speaking children. In G. Keller et al., 1976, 97-142.

Ramirez, A.G. & R.L. Politzer. The acquisition of English and the maintenance of Spanish in a bilingual education program. *TESOL Quarterly*, 1975, *9*, 113-124.

Singh, S. & J. Lynch (eds.). *Diagnostic Procedures in Hearing, Language, and Speech*. Baltimore, Md.: University Park Press, in press.

Swain, M. *Bilingualism as a First Language*. Doctoral Dissertation, University of California at Irvine, 1972.

Waggoner, D. Results of the survey of languages supplement to the July 1975 current population survey. Paper presented at the Fifth Annual International Bilingual Bicultural Education Conference, San Antonio, Texas, 1976.

Using Cloze Procedure as an Overall Language Proficiency Test

Kenneth G. Aitken

Cloze tests are valid, reliable second language proficiency tests. This paper discusses the construction, administration, scoring and interpretation of cloze tests of overall language proficiency. Other uses of the cloze in ESL are mentioned. Finally an explanation of the cognitive processes in doing cloze tasks is offered.

Teachers of English as a second language are beginning to recognize the value of overall language proficiency tests in their programs. One such test of global proficiency is the cloze test, a type of fill-in-the-blank test constructed by randomly deleting words from a prose passage, and replacing them with blanks. Cloze tests appear to be valid, reliable language proficiency tests that can easily be constructed and used by any ESL teacher (Aitken 1975; Stubbs and Tucker 1974; and Oller 1973). The purpose of this paper is to provide some guidelines on the construction, administration, scoring, and interpretation of cloze tests, as well as to discuss a possible explanation of the cognitive processes involved in taking a cloze test. It is hoped that this discussion of the cloze procedure will be comprehensive enough to allow ESL teachers to use the cloze procedure in their own classrooms.

In order to introduce the reader to the subject I will first discuss some of the uses of cloze tests. Next I will explain some procedures for constructing, administering, scoring and interpreting cloze tests. Finally I will conclude with a brief discussion of one possible explanation made by Oller (1973) of the cognitive processes involved in taking a cloze test.

Mr. Aitken is an instructor at Vancouver Community College, Vancouver, Canada

Background and Uses of Cloze Procedure

Cloze tests, pioneered by Taylor (1953 and 1956), have been used for a multitude of purposes (see Jongsma 1971, Oller 1975, and Robinson 1972, for a review of research). Anderson (1971) used cloze procedure to determine the readability, hence, appropriateness, of materials for an ESL class or individual. Extending this idea, one could use materials currently used in the program that the students wish to enter and construct a criterion-referenced test of the student's ability to cope with that material. Jongsma (1971) reports that cloze procedure is a valid and reliable measure of both specific and general reading comprehension. Recently, considerable interest has been shown in the use of cloze tests as measures of overall language proficiency in foreign language. Darnell (1968), using a cloze test and modified scoring system, obtained a high (.83) correlation with the Test of English as a Foreign Language (TOEFL) total test score. Oller (1972) using two different scoring methods but the same cloze test obtained high correlations of (.75) and (.83) with the UCLA E.S.A. placement Examination. Stubbs and Tucker (1974) also reported a high correlation of (.76) between a cloze test and the English Entrance Examination of the American University of Beirut. There appears to be some support for using the cloze as an overall language proficiency test, as well as a reading comprehension measure.

Construction

Constructing a cloze test is a very simple operation. First select a self-contained passage of approximately 375 words. The type of passage selected will depend on the purpose of the test. A test of specific reading comprehension would require a passage previously read. On the other hand, a test of general comprehension, a readability test, or a proficiency test would necessitate choosing a passage as close to the criterion language style as possible. The teacher must be sensitive to the purpose of the test and the goals of the students and the course. In selecting a cloze passage the teacher must be careful not to place a greater demand on the student's knowledge of the content matter than can reasonably be expected of them. Though judgements of content difficulty are generally subjective, reasonably accurate predictions of language difficulty can be made by applying readability formulas to the passages to be used. These formulas usually predict readability, (i.e., passage difficulty) in terms of the minimum reading grade level of persons who can successfully comprehend the passage. Perhaps the most commonly used readability formula is the one devised by Dale and Chall (1948). Recently, however, I have switched to the grade placement formulas devised by Bormuth (1969) to predict 45 and 55 percent cloze criterion scores (the significance of these will become apparent later). Using either formula, it would probably take less than an hour to count up the various difficulty factors and do the addition and multiplication tasks in the formula, working without a calculator.

Second, systematically delete every seventh word, replacing it with a blank of standardized length, until there are fifty blanks. One might conceivably use any n-th word deletion pattern but it has been found that a less-than-every-fourth word, or more-than-every-tenth word deletion pattern is either unmanageable to take or impractical to construct (MacGinitie 1961). Whatever system is chosen, the passage length should be adjusted in such a way as to accommodate about fifty deletions (Pack 1973). Do not choose the items to be deleted, but maintain a uniform random deletion pattern. Some people have taken issue with this and have recommended that proper nouns, numbers, formulas, and dates be skipped over so as to avoid deleting them, maintaining a random pattern thereafter. Whichever view is adopted, consistency is essential. Generally cloze tests have been constructed with the first and last sentences left intact.

Third, prepare the directions and a sample question in simple language.

Administration

The instructions should be read aloud as the student follows his copy on the test. The following are the instructions I use:

> In the exercise below every seventh word has been taken out and blanks put in their place. Your job will be to guess which word has been left out of each space and write that word in the blank.
>
> Remember these things:
>
> (1) Write only ONE word in each blank.
>
> (2) Try to fill in every blank. Guess if you do not know.
>
> (3) You may leave difficult blanks and come back to them when you have finished.
>
> (4) Spelling mistakes will not be marked wrong.
>
> (5) Write neatly please.
>
> (6) Take as much time as you need. It usually takes about half an hour to do the exercise.

An easy sample item follows. For example:

> "When the big man ___________ driving his big car ______________ the road he met _____________ small man in a ____________car driving up the ____________."

After the students have worked this sample item and asked any questions, the students are allowed to begin the test.

Scoring Procedure

When the test has been completed, marking is relatively simple. There are two ways to mark the cloze test; either by accepting only the exact word that was deleted, or by accepting any contextually acceptable answer. Anderson (1971) found that the two methods yielded the same results and recommended the former method because it was easier. Stubbs and Tucker (1974) suggest that the exact word replacement method is as valid as the other method. They found that with a group of 206 university students in ESL cloze tests scored both ways correlated very highly with each other (r=.97). Oller (1972) however, argues that the any-acceptable-answer method is better for use with ESL students.

I have used both scoring methods with children and adults in ESL classes at the beginning, intermediate and advanced stages of second language study. I currently favor the any-contextually-acceptable-answer method. However, there needs to be more research on marker reliability using this method.

In marking, incorrect spellings should not be penalized as long as the word is recognizable. However, the word must be grammatically correct. The wrong verb tense, for example, would not get credit. Each correct (or acceptable answer is worth one point. Double the total (if you use fifty blanks) for the percentage.

Interpreting Cloze Scores

Before entering a discussion on the interpretation of cloze scores it would be advantageous to discuss what is meant by "reading comprehension" and "overall language proficiency." Exactly what is meant by "reading comprehension" has never been satisfactorily determined; however, there has been a certain amount of agreement among reading specialists. Fries (1963) identifies three layers of language meanings that must be dealt with in reading comprehension.

1. Meaning carried by the lexical items.
2. Meanings carried by the grammatical structures.
3. Socio-cultural meanings.

Harris (1969:59) attempts to answer the question, "What is meant by reading comprehension?" in much the same way and concludes that, in practice, the several abilities needed to comprehend the levels of meaning are mutually dependent. Thonis (1970:164) states that reading comprehension in ESL depends on a knowledge of the meanings of words, phrases and sentences, and of arrangements of words, phrases, and sentences according to the conventions of written English. She suggests that the readers must perceive words, phrases and sentences in print; must have experiences which provide a basis for understanding these words, phrases and sentences; must grasp meanings that are greater than the words themselves; and must use this control of print in many different ways. There is more to reading comprehension than meets the eye.

Language proficiency refers to the facility with which an individual can cope with the communication needs of a given task, or in a given situation (Spolsky 1968). These contexts require the individual to function on all three of Fries' levels of meaning and to draw heavily on his own experience bank of real and vicarious communication experiences. Those who attempt to measure language proficiency recognize that what Harris says about skills in reading comprehension being mutually dependent extends into all aspects of language use. Language proficiency is more than mastery of a specific number of discrete structure points and lexical items.

In the final analysis, since overall language proficiency and reading comprehension both depend on the individual's mastery of meaning on the various levels, as well as on the individual's experience bank, a good reading comprehension test is also good test of overall language proficiency. We will return to this point again later.

In multiple-choice reading tests 75 percent and 90 percent comprehension scores are often used for evaluating reading comprehension. A score of 90 percent or greater puts one on what is referred to as an *independent* level. A score of 75 percent to 89 percent puts one on an *instructional* level. Finally, a score below 75 percent puts one on the *frustration* level. The following descriptions of these three levels are suggested as guides, rather than absolute rules, to assist the teacher in determining the relationship between the student's proficiency and the communication task.

At the *independent* level the student can handle the material independently from the classroom or the teacher. A book with a readability rating at this level may be suitable for leisure reading. This range of scores on an overall language proficiency test would indicate that the student would likely experience successful communication in similar circumstances outside the classroom.

When functioning on the *instructional* level the student can manage to cope with this material with guidance or assistance from an instructor. A book on this readability level would be suitable for a supplemental reading activity after an explanation or discussion to introduce the content. This score range on a language proficiency test would indicate that the student would need some form of supplemental second language instruction at a level depending on his actual score.

At the *frustration* level the student cannot understand the material. A book rated in this score range would be much too difficult to understand. A score in this range on a language proficiency test would indicate a need for some basic, intensive ESL instruction.

A comparison of the results of Anderson's (1971) study of cloze tests in ESL with the results from three studies of cloze tests for native speakers at the 75 percent and 90 percent reference points is presented in Table 1.

TABLE 1

Cloze Test Percentage Scores Comparable to 75% and 90% Criterion Multiple-Choice Scores

	Comparable Cloze Percentages			
Criteria	Bormuth (1967)	Bormuth (1968)	Rankin & Culhane (1969)	Anderson (1971)
75%	38	44	41	44
90%	50	57	61	53

In regard to reference points for these previously mentioned levels, Anderson's results seem to be comparable to Bormuth's 1968 results. However it should be noted that Anderson's range within the instructional level is much narrower than the other studies for native speakers. I suspect that his equivalent for the 90% criteria is too low, and that in fact it should probably be closer to Bormuth's 1968 results. This remains to be verified. Needless to say, decisions about individuals based solely on percentages are not entirely just. Caution is required in these areas.

Let us take a hypothetical case to illustrate the cloze test presentation and interpretation. A group of ESL students wish to take a course in beginning bookkeeping. Our question is, "Do they know enough English to follow the lectures and do the assignments in the workbook used in beginning bookkeeping?" One of the measuring sticks we have chosen to help us answer this question is a cloze test. In selecting a passage to use for the cloze we may choose either from the instructional materials used in the bookkeeping course, or from materials within the difficulty range that students enrolled in the course can cope with. Often school regulations restrict enrollment in courses to those who have attained ninth grade, for example. Should this be the case with beginning bookkeeping, we may be able to learn from the school's reading consultant the average Reading Grade Level of native English speaking students in ninth grade at our school, or in the area. Then we may choose a passage that has a safe content (i.e., subject matter is not too strange to be comprehended when a cloze test is made), but still within the reading ability of native speakers taking the course. If our native speaking ninth graders had a Reading Grade Level of 7.5, we might select a passage with a Reading Grade Level of 6.5 or 7. The higher the Reading Grade Level of the passage is, the lower the cloze scores of the student are. After constructing, administering and marking the test we may find that we have scores ranging from 16 to 60 percent score on all the three levels we discussed previously. We would probably consider that those whose scores are on the "instructional" or "independent" levels know enough English to take the bookkeeping course, if the other proficiency measures support that conclusion. We might even pass students whose scores are between 40 and 45 percent, the upper margin of the frustration level, if we have other evidence to suggest that they can cope with both the linguistic and the communication demands placed on students in beginning bookkeeping.

With practice, selecting appropriate cloze passages and interpreting cloze test results in light of your selection criterion becomes easier.

Redundancy and Language Comprehension

One of the fundamental truths about natural language is that it is redundant. Redundancy is that property of language that allows us to predict missing symbols from the context (Carroll 1964). Highly redundant texts tend to be repetitive and contain relatively little information per symbol. By "information" is meant "the informativeness of the symbols" (Shannon and Weaver 1949). Consider the following sentence with two words deleted:

The____(A)______is going_____(B)_______sink in the quicksand.

Deletion (A) could be filled by almost any word representing a tangible thing that could sink in quicksand. When we learn that the missing word is "boy" we note that in this context "boy" contains a lot of information. However, the word missing in deletion (B) could only be "to." Since we can guess this without too much difficulty, we learn that in this context "to" contains relatively little information that is new. The words "going" and "sink," the environment of "to," contain enough information that, when compared with our own internalized grammar, we can unerringly predict the correct answer.

The greater the redundance in a message, the greater the chance that the message will be understood. That the greater the redundance one *perceives* in a message, the greater the chance one will understand it, logically follows.

Redundancy is only helpful to the extent that the receiver of the message is capable of taking advantage of it. If the receiver does not know the structure or rules of the language, or has no previous experiences to give him a feeling of what is likely to be communicated, redundancy is not very helpful. English to an ESL speaker is only redundant to the extent that his concept of English language and communication strategies matches that of a native speaker (see George 1972 for further discussion).

Redundancy is very helpful as it reduces the possibility of errors and misunderstanding and permits communication where there is interference in the communication channel. Messages can be comprehended even though a portion of them are omitted or distorted. However, if we give these mutilated messages to someone who is not a native speaker and does not know the language well, he will experience considerable difficulty in understanding the message. He requires the full normal redundancy (Spolsky 1971). A cloze test is just such a mutilated message. The correspondence between the ESL student's conception of English redundancy and the actual rules of English redundancy becomes an index of his overall language proficiency in English.

When a student responds to the items on a cloze test he is required to project a word to fill in a blank and complete a sequence on the basis of a complete message. He formulates hypotheses or expectations about the information that is to follow by sampling the information that he has available. By further sampling subsequent sequences, the student confirms or disconfirms these hypotheses. If they are disconfirmed, he rejects or revises his expectation, forming a new hypothesis (Oller 1973).

Kenneth S. Goodman (1971) describes the receptive language processes, that is, those of listening and reading in the same terms:

> ... the efficient language user takes the most direct route and touches the fewest bases necessary to get to his goal. He accomplishes this by sampling, relying on the redundancy of language and his knowledge of linguistic constraints. He *predicts* structures, *tests* them against the semantic context which he builds up from the situation and the ongoing discourse and then *confirms* or disconfirms as he processes further language.
>
> Receptive language processes are cycles of *sampling, predicting, testing* and *confirming*. The language user relies on strategies which yield the most reliable prediction with the minimum use of the information available. (p. 136)

Oller (1973) notes that this process is the "mirror image" of what takes place in productive language use (speaking or writing). When a person is trying to convey an idea in speech, he has certain expectations about what he wants to say. He samples his own output and tests and modifies it as he proceeds in order to meet his own expectations. Many of our ungrammatical utterances are the result of such midstream readjustments in our speech.

Further insight into the notion that receptive and productive language processes are integrated comes from Halle and Stevens (1962) who suggest that a listener recognizes what he hears by comparing it with some internal representation which is generated as required by following the same generative rules that normally are used in speaking. If this is true, then the rules of the language are incorporated into a language theory only once in a generative form. The theory, therefore, is that one is a language *user*, not a speaker or listener alone.

Miller (1964) in summarizing Halle and Stevens' suggestion explains:

> The listener begins with a guess about the input. On that basis he generates an internal matching signal. The first attempt will probably be in error; if so, the mismatch is reported and used as a basis for the next guess which should be closer. This cycle repeats (unconsciously, almost certainly) until a satisfactory (not necessarily correct) match is obtained, at which point the next segment of speech is scanned and matched, etc. The output is not a transformed version of the input; it is the programme that was followed to generate the matching representation. (pp. 30-31)

If these processes in "productive" and "receptive" languages are not really distinct, but actually manifestations of the same underlying competence, as Halle and Stevens (1962), Miller (1964), Goodman (1971) and Oller (1973) imply, then a good general comprehension test of reading would be a good overall language proficiency test. Studies of cloze procedure in first and second language tend to bear this out.

In conclusion, I have found, after having constructed, administered and scored over a thousand cloze tests to ESL students in the last three years, that the cloze procedure is an extremely simple, yet valid language proficiency test. Cloze procedure is not a panacea for all ESL testing problems; however cloze tests yield more "miles per gallon" of sweat spent in test construction than most ESL teachers realize. Overall language proficiency tests like the cloze merit much greater consideration than they are given in preparing ESL proficiency test batteries.

REFERENCES

Aitken, Kenneth G. 1975. Problems in a cloze testing re-examined, *TESL Reporter, 8*, 2.

Anderson, Jonathan. 1971. Selecting a suitable "reader": Procedures for teachers to assess language difficulty, *RELC Journal 2*, 35-42.

Bormuth, John R. 1967. Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading 10*, 291-299.

________________. 1968. Cloze test readability: criterion scores, *Journal of Educational Measurement 5*, 189-196.

________________. 1969. Development of readability analyses. Final Report, Project No. 7-0052, Contract No. OEC-3-7-070052-0326. USOE Bureau of Research, U.S. Department of Health, Education and Welfare, March, 1969.

Carroll, John B. 1964. *Language and thought*. Englewood Cliffs, NJ, P Prentice-Hall.

Dale, Edgar and Jeanne Chall. 1948. A formula for predicting readability, *Educational Research Bulletin 27*, 11-20, 37-54.

Darnell, Donald K. 1968. The development of an English language proficiency test of foreign students, using a clozentropy procedure. ERIC ED 025 039.

Fries, Charles C. 1963. *Linguistics and reading*. New York, Holt Rinehart and Winston.

George, H.V. 1972. *Common errors in language learning*. Rowley, Mass., Newbury House.

Goodman, Kenneth S. 1971. Psycholinguistic universals in the reading process. In P. Pimsleur and T. Quinn (eds.) *The psychology of second language learning*. Cambridge, Cambridge University Press, pp. 135-142.

Halle, M. and K.N. Stevens. 1962. Speech recognition: a model and a program for research, *IRE Transactions of Information Theory*, IT-8, 155-159.

Harris, David P. 1969. *Testing English as a second language*. Toronto, McGraw-Hill.

Jongsma, Eugene, 1971. The cloze procedure: a survey of the research. Bloomington, Indiana University School of Education (Occasional Papers in Reading).

MacGinitie, Walter H. 1961. Contextual constraint in English prose paragraphs, *Journal of Psychology, 51*, 1121-1130.

Miller, George A. 1964. The psycholinguists, *Encounter 23*, 29-37.

Oller, John W., Jr. 1972. Scoring methods, and difficulty levels for cloze tests of ESL proficiency, *Modern Language Journal 56*, 151-158.

___________________ 1973. Cloze tests of second language proficiency and what they measure, *Language Learning 23*, 105-118.

___________________ 1975. *Research with cloze procedure in measuring the proficiency of non-native speakers of English: an annotated bibliography.* ERIC ED 104 154.

Pack, Alice C. 1973. Cloze testing and procedure, *TESL Reporter, 6*, 1-2.

Rankin, Earl and Joseph Culhane. 1969. Comparable cloze and multiple-choice comprehension test scores, *Journal of Reading 13*, 193-198.

Robinson, Richard D. 1972. *An introduction to the cloze procedure.* Newark, Delaware, International Reading Association.

Shannon, C. and W. Weaver, 1949. *The mathematical theory of communication.* Urbana, University of Illinois Press.

Spolsky, Bernard. 1968. Language testing - the problem of validation, *TESOL Quarterly 2*, 88-94.

___________________ 1971. Reduced redundancy as a language testing tool. In G.E. Perren and J.L.M. Trim (eds.). *Applications of linguistics*, Cambridge University Press, pp. 383-390.

Stubbs, Joseph Bartow and G. Richard Tucker. 1974. The cloze test as a measure of English proficiency. *Modern Language Journal 58*, 239-241.

Taylor, Wilson L. 1953. Cloze procedure: a new tool for measuring readability, *Journalism Quarterly 30*, 414-438.

___________________ 1956. Recent developments in the use of the cloze procedure, *Journalism Quarterly 33*, 42-48, 99.

Thonis, Eleanor W. 1970. *Teaching reading to non-English speakers.* New York, MacMillan.

IQ TESTS FOR MINORITY CHILDREN AND A PIAGETIAN-BASED COMPUTERIZED INFORMATION SYSTEM AS AN ALTERNATIVE

Edward A. DeAvila

Traditional tests of intelligence are inappropriate for minority children, particularly children of non-English speaking backgrounds. Such diverse groups as the National Education Association, popular press, courts, civil rights organizations, state and federal agencies and school psychologists have all pointed to the failure of the test publishing industry to fully consider the cultural and linguistic differences of minority children when constructing, publishing and selling tests.

Publishers have responded to this criticism by:

1) translating existing intelligence tests for non-English speaking children;
2) adjusting norms for ethnic sub-groups;
3) attempting to construct culture-free tests.

There are distinct problems with each of these approaches. In addition, there are problems concerning the basic validity and utility of information produced by IQ tests. A discussion of inadequacy of the response of test publishers and a presentation of several other issues follows.

Translations

Translating existing intelligence tests for non-English speaking children often creates more problems than it solves. Regional differences within a language make it almost impossible to use a single translation. Thus, while the word "tostón" refers to a quarter or half dollar for a Chicano child, for a Puerto Rican it refers to a squashed section of a fried banana. Monolingual translations are also inappropriate because the language familiar to non-English speaking children is often a combination of two languages as in the case of "pocho" or "tex-mex." Furthermore, many non-English speaking children have never learned to read in their spoken language. One finds many examples of tests written in Spanish being given to Chicano children who may *speak* Spanish but have had no prior instruction in *reading* Spanish.

Another problem in translating tests is that the direct translation of a word or phrase in one language may result in a word which is not used with the same frequency or have the same potency in the second language. For example, the word *pet* is a common word in English. Its Spanish equivalent, *animal doméstico*, is almost never used. Also, translating a word from one language to another can vastly alter its meaning, like the wide variety of seemingly harmless English words which translate into Spanish swear words or "palabras verdes." Thus, translating a large egg into a "huevon" may satisfy grammatical requirements and seem harmless to a translator, but it has a more earthy connotation for Chicanos and Puerto Ricans.

Edward A. DeAvila is the author of several publications concerning tests and their applications. Earlier versions of this article were prepared while he was a Research Director of the Multilingual Assessment Program in California.

Ethnic Norms

The second major response of the testing industry to criticism has been to establish or propose re-establishment of regional and ethnic norms; in other words, to "compensate" minority children for their "deprived backgrounds." Not only will such a practice lead to lower expectations for minorities (which, in turn, lowers children's aspirations to succeed), it is as shortsighted as awarding Chicano children extra points because "they speak a little Spanish." Ethnic norms take no account of the complex reasons why minority children on average score lower than Anglos on IQ, and they are potentially dangerous because they provide a basis for invidious comparisons between different racial groups. The tendency is to assume that lower scores are ultimately indicative of lower potential, thereby contributing to the self-fulfilling prophecy of lower expectations for minorities as well as reinforcing the genetic inferiority argument advanced by Jensen (1971) and others.

Furthermore, if test publishers and users are willing to consider the establishment of ethnic norms, similar arguments could be made for the establishment of norms based on sex differences as well. Considering sex and ethnicity would require an almost infinite set of different norm tables in order to account for all of the different ethnic subgroups in the United States. From the practical point alone this leads to an absurdity. Finally, the establishment of ethnic norms assumes that the groups are ethnically homogeneous with little or no cross-over or intermarriage. One might wonder what publishers would propose to do with a set of male/female twins who had a Mexican father and a Hungarian mother?

Culture-Free Tests

Another way in which the testing industry has responded to criticism of conventional IQ tests has been to create "culture-free" tests. Such tests are difficult, if not impossible, to construct. In tests of mental ability an attempt is made to determine the ability of a child to manipulate certain elements of a problem into a predetermined solution. But if all or some of the elements are not equally familiar to the child, the test is unfairly biased. The influence of culture on conventional IQ test items is subtle in some cases, blatant in others. But the fact remains that in a large number of traditional IQ tests, the items are measuring something other than that for which they were designed. Items particularly influenced by cultural factors fall into the following general classifications:

Socialization. Items of this type are couched in such a way as to actually be measures of the child's family value system. The referent system is, of course, the dominant Anglo-American middleclass. The confounding effects of this problem are particularly evident in the "Comprehension" scale of the Weschler Intelligence Scale for Children (Wisc) where children were asked such questions as:

> "What is the thing to do if you lose one of your friend's balls? or
> "What is the thing to do if a fellow much smaller than yourself starts a fight?"

Allowing for the stilted manner in which the questions are phrased and assuming that the child knows all of the vocabulary, it still seems perfectly obvious that this type of question has little or nothing to do with a child's ability to process, manipulate and/or code information; the answers depend almost exclusively on whether a child has been socialized under the particular ethical system implied by the question.

<ins>Productivity</ins> <ins>or</ins> <ins>Level of Aspiration</ins>. Many tests confound what they hope to measure with a measure of productivity or level of aspiration. For example, in a large number of tests the child who produces the largest number of responses is rewarded while the child who stops responding after only a few attempts is punished by receiving a lower score. Thus, in the Draw-A-Man test the child who produces the more elaborate figure stands the better chance of receiving the higher score, the assumption being that all subjects will produce as many responses as they are able--they all have the same level of aspiration.

Timed tests and "endurance" tests also fall prey to a confusion between the measurement of ability and the measurement of aspiration. In timed tests, which constitute the majority of published group tests, the children are asked to work quickly, quietly and efficiently without regard for the child who is simply not in a hurry and not particularly motivated to be so. The endurance test, for the purpose of boosting statistical reliability, requires that the child answer a large number of questions which vary little in content. This problem is particularly evident in group tests such as the Lorge-Thorndike and the California Test Bureau series. Similar to the "endurance" test is the test where items are sequenced in order of increasing difficulty. This design feature characterizes most of the standardized tests. In this situation the child is forced--by design--to encounter increasing levels of failure and frustration. In the case of the child who starts out fearfully, as do most minority children, the first indication of failure or difficulty is enough to discourage him or her from continuing and one finds the child "staring blankly off into space."

<ins>Experience</ins> <ins>or</ins> <ins>Specific</ins> <ins>Learning</ins>. In tests which require answers of fact, there is an implicit assumption that all the children taking the test will have had an even chance, more or less, of having been exposed to the facts being tested. The spuriousness of this assumption is witnessed by any number of examples where children are asked questions of vocabulary. In such instances, it is impossible to determine whether a minority child has missed a test item because he lacks the capacity to understand a given word or because he simply has never been exposed to the word. <ins>Nitroglycerin</ins> (in the WISC), <ins>fire</ins> <ins>hydrant</ins> (in the Betty Caldwell Preschool Inventory and Peabody Picture Vocabulary Test) or <ins>crevice</ins> (in the Otis-Lennon) are terms unlikely to be heard or used in the home of the average low SES or minority child.

The WISC, perhaps the most widely used individually administered intelligence test in the world, is replete with examples of the importance of specific experience or test results. Take the WISC "General Information" item, "In what kind of store do we buy sugar?" If a child lives in an Indian reservation, he might buy his sugar in a drug store or a trading post.

A Chicano child might reply to that question, "at the Chinitos," a small variety shop owned by a Chinese family. Yet these are not acceptable responses for credit in the WISC. Consider another WISC item, "Where is Chile?" What if a child eats chile, as a Chicano child might. Or consider the item, "What is the thing to do if you lose one of your friend's balls?" The acceptable WISC responses are "Give him one of mine. . .try to get it back or replace it. . .or try to find it." A probable reaction of a barrio or inner city child (depending on the size of the other child) would be avoidance, to escape the child's anger. That is not to say that these kinds of experience-based items are only found in the WISC scale. They are examples of these kinds of biases found in almost all tests of intelligence.

The Validity and Utility of the IQ Score

The basic justification behind the use of the IQ score is that it statistically predicts mental retardation and low achievement. In fact, the IQ test is the sine qua non for screening children suspected of mental retardation. Mercer (1971) found that of those persons who would have been labelled as mentally retarded if their classification depended solely on test scores, a full 84% had completed 8 grades or more in school, 83% had held a job, 80% were financially independent or a housewife, and almost 100% were able to do their own shopping and to travel alone. In other words, it is probable that even at the task for which experts agree the IQ test is best suited--screening for mental retardation--the IQ measure has a dubious real life validity.

The IQ test is also considered by many educators and politicians to be a useful instrument for teachers--for discovering unnoticed learning problems or intellectual strengths, for example. Indeed, many states mandate that districts administer IQ tests several times in a child's succession through the school system. But does the result really help the teacher? Let us take a typical example. A teacher suspects a child of having a severe learning disability. She asks the school or district psychologist to test the child. The psychologist gives the child the WISC in which he scores, let us say, an IQ of 87. This psychologist happens to be extremely conscientious, so he devotes the next few days to writing up an extensive report of his impressions of the child's performance and potential. The child comes from a poor background, the psychologist writes. He has many sibs and is low in the hierarchy with regard to getting attention. After much reviewing the child's school file and his own notes, the psychologist writes a report and hands it to the teacher. The teacher responds in surprise, "But I knew all that. What I want to know is what to do; how can I teach this child." Thus, in most cases neither the psychologist nor the teacher are any wiser despite considerable time and expense administering and evaluating the IQ test. In this situation the psychologist often finds him or herself in the middle of the delicate quasi political-social balance between the classroom teacher, the principal, the district administration, the child and his family.

While few psychologists would agree with the notion that educational decisions affecting the life of the child should be made exclusively on the basis of a single full scale IQ score , the fact nevertheless remains that these decisions are made by educators who, through personal fiat supported by state mandate, ignore both the individual sub-scale profiles as well as psychologists' admonitions, simply for the sake of practical expediency. The result is, of course, a form of default institutional racism.

Informational Needs Within The Educational System and the IQ Tests

Much of the controversy surrounding IQ tests and minority children focuses on whether the IQ model is a valid one. A more practical, and less abstract issue which needs to be answered concerns the general utility of the information produced by the test, i.e., what can one do with the information provided by such a test? In order to answer this question we must consider who is asking the question and for what purpose. It should be apparent that there are qualitative differences concerning the type of information needed within the educational system. These differences depend on the source of the need. To a large extent, much of the confusion surrounding the issue of whether to test stems from a failure to consider these differences. A consideration of the qualitative difference between these different needs will hopefully lend some clarity to the general controversy as well as serve to introduce a procedure that attempts to meet some of the specific needs of each.

There are three general levels of organization within the educational system that require information traditionally obtained through IQ testing. These three may be described as (1) the funding level which involves educational agencies supra-ordinate to the local school districts; (2) the local level which consists of both district personnel and building principals and finally; (3) the classroom teacher, paraprofessional and parent. Whether or not the IQ score information can serve any ultimate useful function will depend on the particular and peculiar needs of each of these three groups. It is unfortunate that historically all three attempt to use the IQ score for radically different and often conflicting purposes.

Consider first the supra-ordinate funding agencies such as state and federal departments of education. For purposes of determining the allocation of funds, these agencies require information concerning educational and program needs at the school district level. While one source of information they generally use are IQ measures, one wonders whether the information produced by IQ tests adds anything of value to needs assessment. More appropriate "needs assessment" procedures, it would seem, would be restricted to the assessment of whether specific educational programs in areas of reading, arithmetic or the like are needed rather than attempting to infer specific need from omnibus assessments based on so poorly understood a concept as IQ.

The supra-ordinate agencies have a second related informational need. In contrast to needs assessments which very often can be conducted through examinations of school attendance records, age-grade placement patterns and achievement data in the broad sense, funding agencies need to know about the effectiveness of particular programs.

Due to "accountability" and "evaluation/audit" requirements, the agencies chiefly responsible for the allocation of funds have mandated that testing be conducted at the child level as a means of collecting information which can be used to evaluate educational programs. In actuality, program evaluations can be made through a variety of means or procedures, none of which necessarily has anything to do with IQ or summary-score types of measures discussed above. For example, reasonable assessment of program effectiveness can be made through the collection of nominal data at various levels of school and community organizations. Thus, administrators, teachers, parents and children can be interviewed as to their perceptions of the effectiveness of the different processes of the program. On the other hand, more product-oriented assessments, at the child level, can be made according to the specific educational objective of the program. The reporting of these latter data can be made on the basis of group change scores, without reference to individual scores, eliminating the potential dangers inherent in individual scores. Finally, all of the various data can be integrated to produce a report which describes the program and weights the relative importance of each aspect of the program in recommending changes in various parts of the total operation.

It is interesting to note that there is a paradoxical element in using IQ scores to evaluate program-related change. The paradox stems from the fact that the IQ model is based on the notion that intelligence is static and hence not subject to change. What this means is that educators have been using IQ scores to evaluate change which can not occur according to the IQ model. The IQ model, which negates time or age through the division of mental age by chronological age, is by definition a static model and, therefore, inappropriate for measuring program-related change which must take place over time.

At the local level, school district personnel need information as to the needs of children and the effectiveness of individual programs in the same way as do the supra-ordinate agencies. However, needs assessments are usually conducted at the state level and school officers more often than not use the state-provided information as a statement of need, preferring not to conduct detailed expensive in-house initiated research.

Ideally, the evaluation of individual programs should center around the collection of data which follows directly from the particular program objectives and activities. However, it is often the case that instruments of evaluation have very little to do with actual programs. Rather than program-specific measures, very often IQ or nationally normed achievement tests are used. Since IQ and other such scores provide precious little in the way of information about the effectiveness of individual programs and program components, there results great confusion in interpreting program evaluations. For example, in comparison with an appropriate "control group" suppose that children who participated in a special program on arithmetic computation had received higher scores on a group IQ score than they had the year before. With this type of evaluation officials would be hard put to say exactly which aspects of the program had led to the improved IQ scores since the connection between the two remains either ill-defined or nonexistent. However, in contrast, a test (administered before and after a specific program) which sampled from each of the different aspects of the program would provide specific information on which aspects of the program led to improved scores and which did not. In other words using IQ scores to evaluate programs not specifically designed for the purpose of evaluating IQ scores is a little like using a thermometer to record height changes in a given period of time.

Unfortunately, the last to be considered in the educational hierarchy is the classroom teacher and what she needs in order to assist the learner. What can the teacher do with an IQ score? The limited utility of the IQ or for that matter any other summary score whether it be reading level according to grade equivalent score, grade point average, percentile rank or the like has already been discussed. Perhaps the fact that to a large extent the tests do not provide the teacher with much usable information is historically related to the fact that psychological tests were initially developed for purposes not functionally related to the present daily needs of the classroom teacher. It is also of some interest to note that even to the present day it is not the classroom teachers who either design the format of the information produced by the testing or, select the specific test or tests which are to be used.

There is still another way in which the use of IQ scores in the classroom might be questioned. Given it is possible to avoid the problems associated with adopting tests which employ the IQ model there is a final question at the teacher-level concerning the ultimate functional use of the IQ score. If it is given that the IQ score is an accurate index of the children in a classroom, then what? What a teacher is asked to do is to discriminate between scores in designing curriculum for each and every child on the basis of information which she can not discriminate. What would a teacher do differently for a child with an IQ of 92 that she would or would not do for a child with an IQ of 100? One might agree that these two scores are functionally equivalent because they are both within the normal range (i.e., within one standard deviation of the mean). However, what about the case of a child who has an IQ score of 84 which is approximately a standard deviation below the mean and in some states considered as possibly falling into "EMR," "retarded" or "slow learner" categories? In the one case where the comparison is made between 92 and 100 there would be no difference in educational recommendation, however, in the other case where scores of 92 and 84 are compared, very different educational recommendations would be made. In both cases there is the same absolute difference in scores.

The point is that we have been basing educational decisions on IQ tests which often produce questionable results. In many cases, the same criticisms apply to achievement tests which provide collapsed scores. Functionally, what programmatic differences can the teacher infer from scores provided in grade equivalent terms. What are the different educational decisions to be drawn from reading grade equivalency scores of 3.2 versus 3.6 and 3.6 versus 4.0? In the case of both the IQ score and the collapsed achievement score, the teacher, who is ultimately the person held "accountable," is simply not being provided with enough information upon which to base sound daily educational decisions. The teacher is thus by default forced to make arbitrary decisions which in most cases can not be defended. The fact that the tests have "predictive validity" in that they forecast failure on the part of Spanish-speaking or other minorities and success for the majority only speaks to the inappropriateness of the curriculum and to the failure of the American educational system to provide education for many of its children.

PAPI: A Piagetian-Based Alternative

In the following, a computer-based model will be described which was designed with many of the above concerns in mind. This model has been successfully tested in a variety of ways in three Southwestern and two Western states. The discussion provided below will outline the elements of the model. More substantive discussion concerning the psychometric features of the measures the model employs are subjects of several other articles currently in the journals or being prepared for publication.

The Program Assessment Pupil Instruction (PAPI) system uses a number of Piagetian-based measures to generate two basic types of information by means of a centralized computer data-processing program.

The first type of information provided by the PAPI system is statistical in nature and is meant for program evaluation at the funding, administrative and district levels. The second type of information is oriented towards the needs of the teacher and consists of recommended activities specifically designed for each child in the classroom.

In its present form, PAPI uses four Piagetian-based measures developed jointly and individually by Pascual-Leone and DeAvila. Two types of measures are used. The first is used as a base measure of achievement where information is collected which follows from classroom activity. The second type of measure attempts to provide an index of the child's current level of development and involves three different tests.

In the following section, each of the four tests in the battery is described. The first to be described is the measure employed as the basic achievement index. The following three are used to provide an index of level of development. The description of the tests is followed by a more detailed description of the PAPI system.

1. CARTOON CONSERVATION SCALES (CCS)

Several measures of Piaget's conservation tasks are assessed by means of a cartoon format developed by DeAvila, Struthers and Randall (1969). In DeAvila's procedure, three horizontal cartoons are presented in which two children are discussing the task. In the first frame an equality is established between two objects according to the dimension being tapped (i.e., number, length, substance, etc.). In the second frame an identity transformation is depicted and in the third frame the question of equality is asked. On the right side of the panel three possible answers are presented in a vertical order. The three alternatives which show the characters responding to the question are randomly ordered as to correctness in order to avoid position effects. Similarly, within a scale wording is altered from panel to panel in order to avoid the possible effects of acquiescence. Background on the conservation scales and an illustration of the dialogue from each scale are presented below.

Struthers and De Avila (1967) and DeAvila, et al (1969) and others have validated the CCS procedure in a variety of ways. In one study DeAvila, et al (1969) tested thirty male and female first grade subjects using both group and individual clinical methods. Statistically significant correlations were obtained across the two methods of assessment (r=0.663) The scales also possessed a high degree of internal consistency as shown by an examination of factor analytic structure and reliability indices.

In its current form the CCS consists of thirty cartoon panels; six examples of five tasks. The panels are presented to the subjects and the story line is read and elaborated upon in order to facilitate understanding of the question. The subject's task is to mark the one (alternative) "that makes the story true."

2. WATER LEVEL TASK (WLT)

The conservation of the horizontality of water described by Piaget and Inhelder (1948), and used by Smedslund (1963), Dodwell (1963), Rebelsky (1964), Beilin, Kagan and Rabinowitz (1966) and Pascual-Leone (in press). This task involves the subject being able to break perceptual set to recognize that no matter what angle a bottle placed on a flat horizontal surface is viewed from, the water level will always be parallel to the horizontal surface. A more complete description of the relative parameters of this type of task can be found in the semantic-pragmatic analysis of Pascual-Leone (in press).

A special version of Pascual-Leone's group tests developed by Pascual-Leone and DeAvila is used. Subjects are presented with individual booklets which contain five horizontal or vertical two-dimensional bottles, eight two-dimensional-tilted bottles and four three-dimensional bottles, two of which are also tilted. The subject is asked to draw a line where the top of the water would be if the bottle were half full and then to place an "X" in the part that contained the water.

3. FIGURAL INTERSECTIONS TEST (FIT)

The figural intersection test is a group administered paper-and-pencil test in which subjects are required to place a dot in the intersection space of a varying number of geometrical figures. It was developed by Pascual-Leone and constitutes a figural analogue to Piaget's work on intersection of classes. In a series of unpublished studies, Pascual-Leone has shown the test to have a high degree of internal consistency (split-half reliability = .89) as well as being significantly related to tests of similar logical structure (Pascual-Leone and Smith, 1968). For example, it has shown a high correlation with the WLT described above. Combined with the WLT, and ST described below, in the present context, it is taken as an index of developmental level. This relationship has been previously found in a series of unpublished studies by Pascual-Leone and Parkinson at York University, and De Avila (1971).

4. SERIAL TASK (ST)

The serial task (DeAvila 1971) is a short term memory task which is individually administered in two phases. First, subjects are pre-exposed to the stimulus materials used in a second testing phase. In the pre-exposure or pre-training phase, each subject is shown a series of 10 different 35mm color slide transparencies of pictures depicting a donkey, house, airplane, etc. Subjects sit facing a screen situated on a wall six feet away. The 10 illustrations are presented by means of a Kodak 650 carousel slide projector. To introduce the task, each subject is shown each figure and asked to give its name and color (i.e., "a yellow hat"). Following this initial introductory phase and after the subject is able to correctly identify each figure ten times when presented in rapid random succession, the testing phase takes place.

The test phase is conducted in a "free recall" manner where, without any prior knowledge of the length of a list, the subject is asked to reproduce the list ignoring the order in which the individual items are presented. Subjects are shown a series of individually presented figures terminated by a blank side, and asked to tell the experimentor what they have seen. The exposure time for each individual slide is .750 msec. There is no requirement that the sequence of the presentation be maintained, or that the subject respond within a specified period of time, or produce a predetermined number of responses. The child is simply asked to reproduce what he has seen using whatever labels are convenient.

There are seven sets of figures presented to each subject. These seven sets vary as to the number of stimuli within a series. The number of figures presented within a series, as well as the individual figures, are randomly varied. Finally, each illustration is presented no more than once within a series.

The PAPI system uses these four measures to produce several types of information each of which has a particular function. First, psychometric information on each test is produced as a means of assuring that the tests used in any given case have satisfactory levels of reliability and validity. This information is of primary importance to the researcher or program evaluator as a check against various types of problems associated with test administration, sampling procedures and so on. The second type of information produced by the PAPI system can be described as developmental-normative. These data provide normative information regarding the current developmental levels of the entire sample. Moreover, these data can also be stored to serve as base-line data which can be compared across temporal (as in a pre-post design) or situational/program dimensions. In summary, these two types of information, the psychometric and the normative-developmental data, are geared to the needs of program evaluation and needs assessment at the administrative levels.

The third type of information generated by the PAPI system consists of lists of recommended classroom activities for each child tested. These activities are geared to providing children with experience in specific educational activities thought to be important in the overall educational/program objectives.

In generating the suggested classroom activities for each child, four factors are taken into account:

1) The child's achievement level with respect to the concepts measured by the CCS.

2) The child's developmental level as measured by the WLT, FIT and ST tests.

3) The achievement level of the child's referent group with respect to the concepts measured by the CCS.

4) The developmental level of the child's referent group as measured by the WLT, FIT and ST tests.

An example of one computer printout which lists specific classroom activities is provided below. Using the four factors listed above, the following example would indicate that:

1. The child's overall developmental test profile is similar to children of the referent group who have already acquired the concept.

2. The child does not understand the concept.

It is then inferred that a specific set of activities are developmentally/educationally appropriate with respect to the concept of number as tested by the CCS. For example, this conjunction of conditions would lead to the following set of recommended classroom activities:

```
                    CONCEPT OF NUMBER

       THIS CHILD DOES NOT HAVE A CONCEPT OF NUMBER,
       HOWEVER, HE IS READY TO LEARN. TRAIN HIM BY
       GOING THROUGH THIS SEQUENCE, USING CONCRETE
       ACTUAL OBJECTS ONLY.

       A.  SORTING AND CLASSIFYING OBJECTS BY
           1.  COLOR
           2.  TEXTURE
           3.  SIZE
           4.  WEIGHT
           5.  NUMBER
           6.  WHATEVER OTHER ATTRIBUTES YOU CAN THINK OF

       B.  COUNTING ACTIVITIES, SUCH THAT THE CHILD LEARNS...
           1.  SEQUENCE, NUMBERS INCREASE BY ONE EACH TIME
           2.  INVARIANCE
               A.  PATTERN, 5 OBJECTS BUNCHED AND 5 OBJECTS
                   IN A ROW ARE BOTH 5

               B.  OBJECTS, 5 ELEPHANTS AND 5 FLIES ARE BOTH 5

           3.  SYMBOLS, LEARN THE ARBITRARY NAMES OF THE NUMERALS
```

C. CONCEPT AND LANGUAGE OF EQUALITIES AND INEQUALITIES
    1. > . WHICH HAS MORE, IS BIGGER, IS LONGER, ETC.

    2. < . WHICH HAS LESS, IS SMALLER, IS SHORTER, ETC.

    3. = . WHICH HAS JUST AS MUCH, IS JUST AS BIG, IS JUST AS LONG, ETC.

The above is an example of what is sent directly to the teacher. It can also be sent to the home so that the parents are made fully aware of what the teacher is trying to accomplish with the child. In this way, possibly with some guidance from the teacher, the parent can participate in the child's education while supporting the teacher.

It should be noted that the PAPI system is designed so that a child's peer or referent group can be designed according to any number of nominal descriptors such as grade, sex or ethnic group. Thus far, only chronological age has been used because of the importance of the age variable in Piaget's theory. It should also be noted that, in the present description of PAPI system, the CCS is treated as an achievement test rather than as a test of development as it was originally designed. Actually, any test of achievement can be used within the system as long as it is linguistically and culturally appropriate and is constructed in such a way that it can be used to generate suggested activities.

In summary, the PAPI system can be used with any child to produce educational program data, or data for program evaluation. The purpose of the present discussion has not been to suggest that the system is in any way a complete educational package. Further development is needed in creating test procedures and in training educators, on how to use this type of system. But given the problems associated with testing, the PAPI system suggests one approach to the issue which speaks to the different needs of people within the educational community.

Summary

The major thesis has been that test publishers and the users of standardized IQ and summary-score achievement tests have failed to consider the problems associated with testing the minority child. And since the results of these tests are used to determine the educational and, by extension, the economic and social future of school-age children, it behooves test publishers and the educational community to take a harder look at the minority child's cultural background. It is the authors' opinion that consideration of these issues leads to the conclusion that the problem of testing cannot be solved by attempts to recreate standardized tests for minority children which are based on old conceptions of intelligence and educational achievement. It is concluded that what is required is a radical change in the whole approach to testing and the generation of entirely different models of education and of testing. There is growing support for this conclusion (for example, see McClelland, 1973). The PAPI system is just one step in an attempt to move in a new direction.

REFERENCES

Beilin, H., Kagan, J. and Rabinowitz, R. Effects of verbal and perceptual training on water level representation. *Child Development*, 1966, *37*, 317-328.

DeAvila, E.A. The use of artificial language in the study of processing capacity, immediate memory and Piaget's cognitive development variable. Unpublished doctoral research, York University, 1971.

DeAvila, E.A., Struthers, J.A. and Randall, D.L. A group measure of the Piagetian concepts of conservation and egocentricity. *Canadian Journal of Behavioural Science*, 1969, *1 (4)*, 263-272

Dodwell, P.C. Children's Understanding of number and related concepts. *Canadian Journal of Psychology*, 1960, *14*, 191-205.

Jensen, A.R. Do schools cheat minority children? *Educational Research*, 1971, *14*, 3-28.

McClelland, D. Testing for competence rather than for "intelligence." *American Psychologist*, 1973, *28*, 1-14.

Mercer, J. The meaning of mental retardation. In Richard Koch and James Dobson (Eds.), *The mentally retarded child and his family: A multidisciplinary handbook*. New York: Brunner/Mazel, Inc., 1971.

Pascual-Leone, J. and Smith, J. The encoding [illegible] decoding of symbols of children. A new experimental paradigm and a neo-Piagetian model. *Journal of Experimental Child Psychology*, 1968, *8*, 328-355.

Piaget, J. and Inhelder, B. *La representation de l'espace chez l'enfant*. Paris: P.U.F., 1948. English translation: *The Child's Conception of Space*. London: Routledge and Kegan Paul, 1956.

Rebelsky, F. Adult perception of the horizontal. *Perceptual and Motor Skills*, 1964, *19*, 371-374.

Smedslund, J. The effect of observation on children's representation of the spatial orientation of a water surface. *Journal of Genetic Psychology*, 1963, *102*, 195-201.

Struthers, J. and DeAvila, E.A. Development of a group measure to assess the extent of pre-logical and pre-casual thinking in primary school age children. Paper presented to National Science Teacher's Association Convention, Detroit, 1967.

LIST OF ABBREVIATIONS AND DEFINITIONS PERTINENT TO THE TEXT

TEFL (EFL)	Teaching English as a Foreign Language...identifies programs of English language instruction in parts of the world where some other language is spoken, e.g., just as the study of Chinese in United States schools is classed as "foreign language" study, so the study of English in France comes under the heading of "foreign language" study. An English language program in Turkey, therefore, would be an example of a TEFL program.
TESL (ESL)	Teaching English as a Second Language...identified programs of English language instruction found in English speaking areas but intended to benefit non-English speaking persons in those areas, e.g., an English language program of instruction in London for Italian speakers living there would be an example of a TESL program; all English language courses for Spanish speakers offered in the New York City public schools are TESL programs. We can now make the following TEFL/TESL distinction: the same English language course for non-English speakers offered in Mayaguez, Puerto Rico and in Brooklyn, New York, would be given two different classifications, the one in Mayaguez would be an example of TEFL, and the one in Brooklyn would be TESL.

TESOD (ESOD) or SESD Teaching English to Speakers of Other Dialects or Standard English as a Second Dialect...identifies programs of General Standard English instruction, generally found in English speaking communities and intended to benefit English speakers who use a non-standard pattern of English, e.g. a course in General Standard American English for cockney English speakers would be an example of TESOD.

TESOL (ESOL) Teaching English to Speakers of Other Languages..is simply a broader classification of convenience to include all three already identified under one heading, i.e., TEFL, TESL, and TESOD are all specialized activities under the heading of TESOL.

These classifications can be conveniently projected to include other language study areas, e.g., S for Spanish, C for Chinese, G for German, P for Portuguese, N for Navajo etc. For Spanish language study programs for example, we can establish the following classifications:

TSFL (SFL) Teaching Spanish as a Foreign Language...for Spanish language programs found in China, Japan, England, Turkey, the United States, etc.

TSSOD (SSOD) Teaching Spanish to Speakers of Other Dialects...as in the case of standard Castilian courses for Andalusian Spanish speakers.

TSSOL (SSOL) Teaching Spanish to Speakers of Other Languages...a convenience classification which would include TSFL, TSSL, and TESSOD.

English as a Second Language is also used to identify the use of English (by common agreement) as a lingua franca in multilingual areas, as is the case in some of the new countries of Africa where the citizenry is composed of numerous different speech communities, and where no satisfactory agreement can be reached on making any one of the native languages the "official" language of the new nation. A temporary and makeshift solution is to make some outside language, be it English or French, or Portuguese, or some other, the language of government and official transactions. This type of solution avoids intense, internal jealousies but commits everyone to having to learn a second language.

Services Provided by the New York State Education Department, Bureau of Bilingual Education

The Bureau of Bilingual Education, with regional offices in Albany, Buffalo, New York City and Long Island, provides services to the State in bilingual education/English as a second language programs. The Bureau's activities, presently involving twenty-two different language groups, include the following:

I. Technical Assistance to school districts in areas such as: on-site teacher training workshops; school-community relations; desegregation planning; and preparation of proposals for funding. Technical assistance is also provided by letter and telephone.

II. Monitoring of ESEA Title VII, ESEA Title I and State Chapter 720 programs in the schools.

III. Publications presently include: Haitian, Chinese, Spanish, Native American, Greek, French, and English as a second language materials, documents on bilingual programs and testing, teachers' handbooks, and a quarterly newsletter.

IV. Media Projects include: a series of public service announcements on bilingual education, video-tapes of conference presentations such as Judith Nine-Curt's presentation on Puerto Rican body language, Spanish translations of several films, and a half-hour video tape on bilingual education.

V. Resource Library (in the Albany office) includes a wide range of materials in bilingual education and English as a second language. Upon request, materials are available on loan to schools throughout the State.

VI. Program Planning and Coordination for bilingual education, including the gathering of statistics demonstrating the need for (and performance of) bilingual education programs.

VII. Bilingual Higher Education services are provided by Bureau of Bilingual Education personnel working in the Grants Administration Office.

VIII. Bilingual Migrant Programs are developed and coordinated for migrant children throughout the State.

The Bureau of Bilingual Education serves thousands of children and adults in all regions of the State. For more information, please call:

Albany Office:	
Coordinating and Central New York	(518) 474-8775
Buffalo Office:	
Western New York	(716) 686-2087
Long Island Office:	
Long Island and Westchester	(516) 997-8700 ext. 370
New York City Office:	
New York City Area	(212) 488-7294
Higher Education Office:	
Statewide	(518) 474-5705